Responsible AI Toolkit for African Startups

DIFA CONSULTANCY

UK International Development
Partnership | Progress | Prosperity

AKILI AI

KICTANet
The Power of Communities

About DIFA Consultancy

DIFA Consultancy is an advisory firm specialising in responsible artificial intelligence (AI), data protection, technology policy and governance. The firm supports businesses, non-profits, and public sector institutions in navigating responsible AI adoption and data privacy obligations through ethical, compliant, and context-responsive strategies and advisory. Based in Kenya and working across Africa and globally, DIFA integrates international best practices with local regulatory insights to deliver tailored advisory services, capacity-building, and applied research and policymaking expertise that advances secure, accountable AI and resilient data and digital ecosystems.

Acknowledgements

Authors:

This Responsible AI Toolkit for Startups in Africa was prepared by:

Ikran Abdirahman,

Founder and Principal Consultant, DIFA Consultancy.

Betsy Muriithi,

Technical AI Consultant.

Natasha Karanja,

Legal and AI Ethics Consultant, DIFA Consultancy.

Special acknowledgments:

We are grateful for the support received from our partners, and to our community of experts for their invaluable contributions throughout the implementation of the project.

Funding and Implementation Partners:

- **Wilfred Oluoch Omondi**, a representative of the British High Commission, the funding partner in this project.
- **Dr. George Musumba,** a representative of the Kenya ICT Action Network (KICTANet), the implementing partner of the AI Challenge Fund.
- **Martin Mbaya,** of Akili AI, for his meaningful support in steering the activity from the conceptualization to implementation stages.

Contributors and Content Reviewers:

- Stephanie Kasaon, DAIvolve Technologies and Action Lab.
- Martin Mbaya, Akili AI.
- Ridwan Oloyede, Tech Hive Advisory Africa.
- Cecil Abungu, ILINA Program.
- Melissa Omino, Florence Ogonjo and the team, Centre for Intellectual Property and Information Technology Law (CIPIT) at Strathmore University.

Project Management and Coordination:

- Faith Nekoya Musumba, Project Coordinator, DIFA Consultancy.
- Kimberly Fiona Okiri, Projects and Events Assistant, DIFA Consultancy.
- Firdaus Maalim, Research Intern, DIFA Consultancy.

Graphics and Cover Design: Dennis Miano

Disclaimer:

This publication is provided for general informational and educational purposes only and does not constitute legal advice or a legal report. The content, including any responsible AI assessment frameworks and regulatory compliance considerations, is not exhaustive and may not apply to all situations, nor does it provide a holistic analysis of all relevant issues. While efforts have been made to ensure accuracy, DIFA Consultancy does not guarantee the completeness of the content.

Published in 2026 by DIFA Legal Consultancy Limited, 4th Ngong Avenue Towers, Nairobi, Kenya. Email: info@difaconsultancy.com.

This Toolkit was developed as part of the British High Commission funded project, “Development of a Responsible AI Toolkit for Startups in Africa, and Capacity Building Training,” in partnership with Akili AI, and with the support of KICTANet.

Table of Contents

Legal and Policy Instruments

Abbreviation / Short Name	Full Term
AU -AI Continental Strategy	African Union Continental Artificial Intelligence Strategy (2024)
AU Digital Strategy for Africa	African Union Digital Transformation Strategy for Africa (2020-2030)
AU Malabo Convention	African Union Convention on Cyber Security and Personal Data Protection (2014)
Botswana DPA	Data Protection Act, 2024 of Botswana (Act No. 18 of 2024)
Copyright Amendment Bill 2017	South Africa Copyright Amendment Bill, 2017
Digital Credit Providers Regulations 2022	Central Bank of Kenya (Digital Credit Providers) Regulations, 2022
Egyptian Copyright Law	Law No. 82 of 2002 on the Protection of Intellectual Property Rights
Egyptian Data Protection Framework	Law No. 151 of 2020 on the Protection of Personal Data (PDPL)
EU AI Act	European Union Artificial Intelligence Act (Regulation (EU) 2024/1689)
EU DSM Directive Article 4	Directive (EU) 2019/790 on Copyright in the Digital Single Market, Article 4
Indigenous Knowledge Act	South Africa Protection, Promotion, Development and Management of Indigenous Knowledge Act, 2019
Protection of Traditional Knowledge and Cultural Expressions Act	Protection of Traditional Knowledge and Cultural Expressions Act, No 33 of 2016 (Kenya)
Kenya AI Strategy	Kenya Artificial Intelligence Strategy 2025-2030
Kenya Data Protection Act 2019	Kenya Data Protection Act, No. 24 of 2019
Kenya's Copyright Act	Kenya Copyright Act No 12 of 2001
Kenya's Industrial Property Act	Kenya's Industrial Property Act No 3 of 2001
NDPA 2023	Nigeria Data Protection Act, 2023
Nigerian Trade Secrets provisions	Nigerian Trade Secrets Legal Framework
POPIA	Protection of Personal Information Act, 2013 (South Africa)
Protection of Traditional Knowledge and Cultural Expressions Act	Kenya Protection of Traditional Knowledge and Cultural Expressions Act, 2016
South Africa's Copyright Act	South Africa Copyright Act, No. 98 of 1978
South Africa's Patents Act	South Africa Patents Act, No. 57 of 1978
Swakopmund Protocol	Swakopmund Protocol on the Protection of Traditional Knowledge and Expressions of Folklore
Standards and Frameworks	
KEBS AI Code of Practice	Kenya Bureau of Standards AI Code of Practice
NIST AI RMF 1.0	National Institute of Standards and Technology Artificial Intelligence Risk Management Framework 1.0
OECD AI Principles	OECD Recommendation on Artificial Intelligence (2019)

Executive Summary

Artificial intelligence (AI) is increasingly becoming a core aspect of growth and a competitive differentiator in many startups in Africa. This broad integration necessitates the adoption and implementation of Responsible AI (RAI) as a strategic startup endeavour for safe AI development and deployment, rather than pushing RAI as a late-stage compliance or regulatory exercise. However, resource constraints and limited technical expertise frequently impede the effective adoption and implementation of RAI principles by these startups.

This Responsible AI Toolkit for African Startups (Toolkit) was developed to close that gap. It provides accessible, actionable guidance tailored to the context of the African startup ecosystem, equipping startup teams with the frameworks needed to develop and deploy AI systems that are innovative yet ethical, and scalable yet trustworthy. Specifically, the Toolkit balances innovation priorities whilst being cognizant of safety needs, ensuring that AI is built and brought to market in ways that are ethical, inclusive, and aligned with long-term business and societal value.

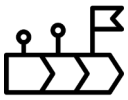
The Toolkit, therefore, aims to simplify and provide comprehensive guidance to startups to navigate the understanding and implementation of RAI in their functions. It empowers startups to balance innovation with safety, incorporate RAI principles into building AI, and mitigate risks related to bias, data protection, and intellectual property concerns. Through the Toolkit, startups also have a roadmap to navigating relevant regulatory landscapes across African markets.

As such, the Toolkit is designed primarily to support startups with limited resources that may lack dedicated expertise or capacity to navigate the complexities of ethical AI while competing with more advanced corporations. At the same time, the Toolkit is equally valuable to more mature or well-resourced startups, offering deeper guidance, frameworks, and best practices that can strengthen existing governance processes and advance RAI practices.

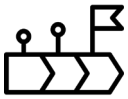
Ultimately, this Toolkit reframes RAI from being perceived only as a compliance cost, into a strategic enabler. It encourages African startups to harness and implement AI not only to solve pressing challenges across the continent, but to do so in a manner that is ethical, sustainable and globally competitive in line with regional and global best practices.

Toolkit Structure

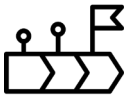
To operationalize these RAI principles, this Toolkit provides a structured roadmap for African startups across three comprehensive modules:



Foundational Principles of Responsible AI: The first module establishes the foundational principles of AI, defining core values such as fairness, transparency and accountability as the bedrock of the innovation process.



Regulatory Compliance and Governance in African AI Landscape: The second module navigates the regulatory compliance and governance landscape specifically within the African AI ecosystem, offering guidance on legal requirements and ethical standards across diverse jurisdictions.



Technical AI Risk Assessment and Mitigation Framework: The final module provides a lifecycle oriented risk framework that enables startups to systematically classify, assess and mitigate technical risks from initial development through to ongoing deployment and oversight.

Beyond the above modules, the Toolkit also provides practical application guidance across the AI lifecycle through the following appendices:

- **Appendix 1** presents a case study which serves as a practical learning exercise, guiding the user in applying RAI principles, conducting AI risk assessments, and identifying relevant legal and regulatory compliance considerations.
- **Appendix 2** provides a detailed assessment checklist and governance toolkit aligned with RAI principles. It also highlights key risk categories that may arise at different stages of the AI lifecycle.
- **Appendix 3** compiles a list of RAI risk assessment and mitigation tools available in the public domain, which are mapped across the AI lifecycle, to assist the user in undertaking identification, evaluation and management of AI-related risks.
- **Appendix 4** outlines a training delivery framework designed to guide and support facilitators using the Toolkit in effectively implementing and delivering the content in a training programme.

Why Responsible AI Matters in the African Startup Ecosystem?

Artificial Intelligence (AI) powers disruptive startups not as a peripheral tool, but as the central product and growth engine. This deep integration makes the principles of **Responsible AI (RAI)** not a corporate afterthought, but a critical component of the foundation. In the startup context, where speed, trust and scalability determine survival, RAI transforms from a risk management cost to a strategic enabler and competitive advantage.[1] While legal frameworks establish a critical compliance floor, they are inherently reactive and minimalistic, failing to address the broader spectrum of ethical, reputational and operational hazards.[2]

Therefore, RAI is not a peripheral compliance checklist or a public relations exercise. It is a core strategic discipline essential for sustainable value creation. The central business imperative is to construct an AI strategy that actively integrates RAI principles, such as fairness, transparency, accountability and safety directly into the innovation lifecycle.[3] This proactive integration serves as the vital mechanism to:

01 **Operationalizable and guided frameworks:** Conceptual goals should be cast into specific, documented strategies. By identifying accountable parties: from engineers to executives, at every stage of the AI lifecycle (conception, development and monitoring), startups ensure that responsibility is assigned to those with the capacity to implement it.[4]

02 **Flexible and iterative implementation:** Given the volatility of emerging markets, RAI frameworks must be flexible enough to adapt to various use cases and organizational settings. Implementation must be iterative, recurring throughout the system's lifecycle as the implementation context or external environment shifts.[5]

For AI startups navigating African markets, this Toolkit serves as a resource for founders to autonomously map their specific area of operations against compliance responsibilities and their localized risk appetites. This Toolkit centralizes critical compliance and AI risk information, while the burden of establishing an informed defensible, contextualized position where proactive risk mapping lies with the startups themselves. By utilizing these hands-on resources to systematically identify pitfalls and match risks proportionately to specific localized obligations, founders ensure their business possesses the institutional maturity to withstand rigorous due diligence and regulatory scrutiny as it expands.

In essence, for the modern African enterprise, this proactive integration serves as the vital mechanism to protect value by safeguarding brand reputation and customer trust. It ensures operational resilience by building governable systems that reduce the risk of unmanageable liabilities, and it enables responsible innovation by providing the guardrails necessary to scale powerful AI aggressively. A startup's commitment to these RAI principles forms the basis and architecture for sustainable and scalable growth, and is not a constraint on innovation.

1 PwC, Responsible AI < https://www.pwc.com/gx/en/services/ai/responsible-ai.html>

2 Alistair Reid, Simon O'Callaghan & Yaya Lu, Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources [2023] Gradient Institute & CSIRO,2

3 Lizzie Short, Building a Responsible AI Framework: 5 Key Principles for Organizations [2025] Harvard University: Professional & Executive Development <https://professional.dce.harvard.edu/blog/building-a-responsible-ai-framework-5-key-principles-for-organizations/#What-Is-Responsible-AI >

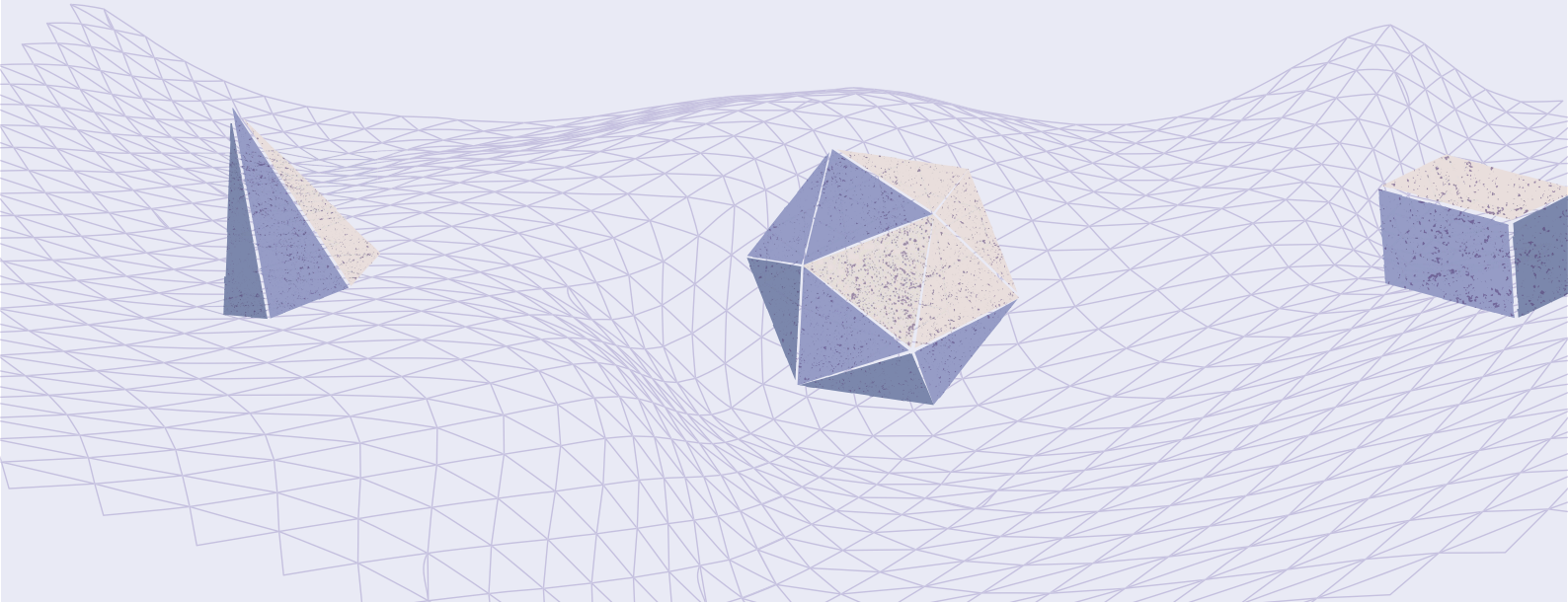
4 ibid

5 ibid

01

Module 1

Foundational Principles of Responsible AI



Objective

RAI is a governance framework that ensures the ethical and lawful deployment of AI.[6] The framework provides organizations with structured guidelines to ensure the proactive management of risks posed by AI while maximizing its positive impact for individuals and society at large.[7] Therefore, through formal frameworks, organizations define their approach to developing, designing and deploying AI that is not only innovative but also trustworthy and socially beneficial.[8] This section establishes the core ethical and legal principles of RAI, providing African startups with the conceptual understanding necessary to anchor their practices. It explains the 'why' behind the rules, connecting global standards to local context and demonstrating how these principles form the bedrock of trustworthy and responsible innovation.

The Global and Continental Architecture of Responsible AI: A Unified Foundation

The convergence of global and continental governance frameworks has established a definitive set of foundational principles that serve as the baseline for operationalizing Responsible AI. These core tenets, ranging from human wellbeing to contestability, have been codified by leading international bodies to ensure that AI development remains aligned with shared ethical values.

The OECD and G20 AI Principles represent a landmark global consensus, emphasizing that AI systems must drive inclusive growth and sustainable development while promoting human, societal and environmental wellbeing. These principles prioritize human centered values and fairness by requiring AI actors to respect the rule of law and human rights throughout the system lifecycle. Crucially, these frameworks highlight transparency and explainability as the technical and procedural basis for contestability, allowing individuals to understand and challenge AI driven outcomes. In parallel, the UNESCO Recommendation on the Ethics of Artificial Intelligence expands this foundation by integrating social wellbeing as a primary objective, and codifying privacy and security as fundamental rights. Complementing these, the IEEE Ethically Aligned Design initiative[9] prioritizes human rights and introduces data agency as a foundational principle to empower individuals with control over their digital identities.

---

6 World Economic Forum, Why we need to care about Responsible AI in the age of the algorithm [2023] < https://www.weforum.org/stories/2023/03/why-businesses-should-commit-to-responsible-ai/>

7 ibid

8 ibid

9 IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, IEEE Global Initiative Releases Treatise on Ethically Aligned Design of Autonomous and Intelligent Systems [2019] <https://globalpolicy.ieee.org/ieee-global-initiative-releases-treatise-on-ethically-aligned-design-of-autonomous-and-intelligent-systems/>

Continentally, the African Union has aligned its AU-AI Continental Strategy with these global baselines while ensuring they are culturally and socio economically relevant. African governance discussions often anchor these principles in local moral frameworks to ensure that fairness and inclusivity respect the interconnectedness of individuals and their communities. Technical and regulatory benchmarks further operationalize these values into measurable requirements. The NIST AI Risk Management Framework defines core characteristics of trustworthy AI, such as safety and resilience, through a lifecycle oriented approach that transforms abstract values into technical requirements. Meanwhile, the EU AI Act translates these principles into mandatory legal requirements, moving beyond voluntary guidelines to enforce transparency and accountability as a standard for market access and operational legitimacy. Collectively, these global and continental instruments demonstrate that human centered values, fairness and privacy are now the recognized foundational pillars of the international AI landscape.

Principles of Responsible AI

The governance of Responsible AI is underpinned by primary foundational principles that transform abstract ethical concepts into measurable operational requirements. These main principles serve as the essential quality metrics for a system's lifecycle. By adopting this collective framework, organizations ensure that their technological developments remain technically robust and socially beneficial, providing the necessary oversight to align autonomous outcomes with human autonomy and the common good.

The following principles constitute the core foundation of Responsible AI:

Fairness

Human-centered values

Accountability

Data privacy and security

Transparency and Explainability

Contestability

Fairness and Non-Discrimination

The application of AI today must be anchored in the foundational principle of fairness , establishing it as the non-negotiable starting point for system design and deployment[10]. Despite its vitality, the principle is not uniformly defined[11]. A coherent understanding of fairness in AI begins by recognizing its multifaceted nature, as interpretations differ across technical specifications, cultural norms, societal values and legal standards[12]. Therefore, a responsive approach to fairness must acknowledge its many interpretations and contextual applications and this same principle extends to related concepts such as equity, impartiality, justice, equality and non-discrimination[13].

10 Markus Kattnig, Allesa Angerschmid, Thomas Reichel & Roman Kern, Assessing trustworthy AI: Technical and legal perspectives of fairness in AI [2024] Computer Law& Security Review, 2

11 ibid

12 ibid

13 ibid

Definitions of Key Concepts:

To foster a shared understanding, it is crucial to clarify these interconnected concepts:

- **Fairness:** In AI, fairness broadly refers to the principle that all individuals are entitled to equal rights and should be treated impartially. It encompasses both distributive justice and socio-relational dimensions, striving for just outcomes by considering individuals' unique social, cultural, and environmental factors.
- **Bias:** This refers to systematic and unfair discrimination in AI systems, often originating from unrepresentative or flawed training data, biased algorithms or human decision-making processes. It can lead to outcomes that disadvantage certain groups or perpetuate existing inequalities.
- **Equality:** Ensures consistent and impartial treatment for all individuals under the law, aiming for equal access and outcomes without differentiation.
- **Equity:** Focuses on achieving just outcomes by distributing resources and opportunities based on nuanced needs and circumstances. It recognizes that certain groups may require tailored resources or support to attain comparable benefits.
- **Impartiality:** Operating objectively, without favoritism or prejudice, ensuring decisions are made based on merit and relevant criteria.
- **Justice:** The overarching moral and legal principle guiding fairness, equality and equity, ensuring rights are respected and wrongs are remedied.
- **Non-discrimination:** The principle prohibiting direct or indirect differentiation based on protected characteristics (e.g., race, sex, disability), ensuring all individuals are treated equally under the law.

a.

Legal Fairness

Examining fairness through a legal lens requires understanding the distinct yet complementary roles of equality and equity. Equality ensures consistent and impartial treatment for all individuals under the law. Equity, in contrast, focuses on achieving just outcomes by distributing resources and opportunities based on nuanced needs and circumstances. Together, these principles provide a complete framework for pursuing substantive justice. This is supported by international instruments such as the Universal Declaration of Human Rights (1948) that emphasizes the equality principle by acknowledging we are equal under the law and are entitled without any discrimination to equal protection of the law.[14] Regionally, we have the African Charter on Human and Peoples' Rights (1986) that provides for equality and non-discrimination as core principles to be observed for the achievement of the legitimate aspirations of the African peoples.[15] Exploring this within the AI context, the use of AI systems may contribute to the restriction of rights or opportunities of individuals that may hinder their right to equality. This is illustrated through examples such as AI automated decision-making or decision support systems that may have the potential to influence how individuals enjoy their rights. The legal implications for AI systems become critically clear. The operation of these systems must be aligned with these established principles to avoid perpetuating or exacerbating existing injustices.

Therefore, for startups building or integrating AI, the legal concept of fairness transcends abstract ethics; it signifies a concrete duty of care. This duty involves proactively designing systems that do not simply process data neutrally but are actively scrutinized and shaped to prevent discriminatory outcomes and promote inclusive access. It requires embedding governance mechanisms, such as bias audits, impact assessments and human review processes into the AI lifecycle. By anchoring their practices in the legal principles of equality and equity, startups not only mitigate regulatory risk but also build trust, foster sustainable innovation and ensure their technologies contribute positively to the societies they serve.

This accountability is increasingly codified in national legislation. Modern data protection laws, which serve as the primary regulatory touchpoint for AI in many jurisdictions, explicitly address automated processing. For instance, statutes such as Kenya's Data Protection Act (2019) and South Africa's Protection of Personal Information Act (POPIA) grant individuals the right not to be subject to decisions based solely on automated processing, including profiling, which significantly affects them.[16] This legal provision empowers individuals to challenge decisions made by AI systems, thereby instituting a crucial check against automated unfairness.[17] Furthermore, these laws often mandate Data Protection Impact Assessments (DPIAs) for high-risk processing activities.[18]

14 Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III)

15 African Charter on Human and Peoples' Rights (adopted 27 June 1981, entered into force 21 October 1986) 21 ILM 58 (1982)

16 Kenya Data Protection Act (No 24 of 2019), s 34 & South Africa Protection of Personal Information Act (No 4 of 2013), s 71

17 ibid

18 Kenya Data Protection Act (No 24 of 2019) s31

In an AI context, high-risk activities include:

- **Large-scale processing of sensitive personal data:** For instance, using biometric data for identification or health data for diagnostic purposes.
- **Automated decision-making with legal or significant effects:** Such as credit scoring, employment decisions or eligibility for public services.
- **Systematic monitoring of publicly accessible areas on a large scale:** For example, facial recognition in public spaces.
- **AI systems that profile individuals extensively:** Especially when this leads to segregation or exclusion of individuals.

For a startup, this means conducting a thorough evaluation of how an AI system might infringe on the right to equality before deployment, assessing potential biases in data and algorithms, and implementing mitigations, a practical application of the principle of equity.

Looking beyond the continent for a comparative perspective, global regulatory trends further solidify these requirements. The European Union's General Data Protection Regulation (GDPR) sets a stringent precedent with its rights to transparency and explanation in automated decision-making.[19] More specifically, the EU AI Act directly legislates based on risk, prohibiting AI systems that pose an unacceptable threat to fundamental rights and imposing strict obligations for high-risk applications.[20] These obligations include fundamental rights impact assessments, bias mitigation and human oversight measures designed to enforce both equality of treatment and equity of outcome. Similarly, the OECD AI Principles, adopted by over 50 countries, establish a global standard that AI systems should respect human rights and democratic values operating with transparency and fairness.[21]

19 General Data Protection Regulation (GDPR) Regulation (EU) 2016/679 [2016] OJ L 119/1

20 EU AI Act Regulation (EU) 2024/1689 [2024] OJ L2024/168

21 OECD, 'Recommendation of the Council on Artificial Intelligence' (adopted 22 May 2019, amended 3 May 2024) OECD/LEGAL/0449

b.

Discriminatory Non-Harm

This principle establishes the essential threshold for the "do no harm" imperative, which is breached through acts of direct or indirect discrimination.[22] It is also violated by discriminatory harassment tied to a protected characteristic.(e.g race, sex, disability) actions that undermine dignity, degrade identity, or foster a hostile, humiliating or offensive environment for individuals.[23]

Therefore, for startups building or integrating AI, the legal concept of fairness transcends abstract ethics; it signifies a concrete duty of care. This duty involves proactively designing systems that do not simply process data neutrally but are actively scrutinized and shaped to prevent discriminatory outcomes and promote inclusive access. It requires embedding governance mechanisms, such as bias audits, impact assessments and human review processes into the AI lifecycle. By anchoring their practices in the legal principles of equality and equity, startups not only mitigate regulatory risk but also build trust, foster sustainable innovation and ensure their technologies contribute positively to the societies they serve.

22 Leslie(n16)

23 ibid

Demonstrated Examples of AI Discrimination:

The ethical implications are starkly illustrated by real-world instances of AI discrimination:

Facial Recognition in African Context: Facial recognition systems, often trained on non-diverse datasets, have shown significantly higher error rates for individuals with darker skin tones, posing serious challenges for people of African descent.[24]

Uber's Algorithmic Management: Studies have revealed how Uber's algorithmic management systems, particularly concerning dynamic pricing and task assignment, can impact worker autonomy, fairness and long-term work relations. This has raised concerns about decreased pay, increased unpredictability and enhanced inequality among drivers, highlighting the power imbalance between algorithmic systems and workers.[25]

By anchoring their practices in the legal principles of equality and equity, and proactively addressing data representativeness and bias, startups not only mitigate regulatory risk but also build trust, foster sustainable innovation, and ensure their technologies contribute positively to the societies they serve.

24 Notice Pasipamire & Abton Muroyiwa, Navigating algorithm bias in AI: ensuring fairness and trust in Africa [2024] Front.Res. Metr, Anal, Vol 9

25 Reuben Binns et al, 'Not Even Nice Work If You Can Get It; A Longitudinal Study of Uber's Algorithmic Pay and Pricing' in The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)

C.

Technical Fairness in AI Fundamentals

When it comes to technical fairness in AI, unfair outcomes are rarely caused by "the model alone," but by interactions between social and technical factors that include (i) organizational goals and incentives, (ii) data and modelling choices, and (iii) how outputs are implemented in real decisions and institutions. Bias and discrimination can enter at many points, from agenda setting and data extraction to model development and deployment.[26] Legal fairness distinguishes between direct and indirect discrimination.[27]

On the one hand, direct discrimination (disparate treatment) occurs when individuals are treated adversely because of protected characteristics (e.g., race, sex, disability). In AI systems, direct discrimination can happen if protected characteristics (or explicit group membership signals) are used directly in rules, features, thresholds, or segment-specific models, producing intentionally different treatment for similarly situated people.

Direct Discrimination	Key Insight/Actionable Steps
Example: A hiring tool that downgrades applicants who indicate maternity leave gaps or disability accommodations leading to systematically lower hiring outcomes for women or persons with disabilities.	• Direct discrimination is based on protected characteristics (gender, disability) which carries explicit penalization in the model. • Implement governance mechanisms to prevent direct discrimination, including bias audits, impact assessments, and human review processes integrated into the AI lifecycle.

On the other hand, indirect discrimination (disparate impact) occurs when seemingly neutral criteria, policies, or practices disproportionately disadvantage protected groups. In AI systems, this is often the dominant risk because it can occur without explicitly using protected attributes: proxy variables (e.g., location, school, device type, language) can correlate with protected characteristics and recreate discriminatory patterns ("proxy discrimination"), which is often hard to detect.

26 Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ. Burr, C., Aitken, M., Katell, M., Fischer, C., Wong, J., and Kherroubi Garcia, I. (2023). AI Fairness in Practice. The Alan Turing Institute.

27 Marcus (n10)

Indirect (proxy) discrimination	Key Insight/Actionable Steps
Example: A digital lender model "learns" that people with budget phones are risky because many lack formal credit history in the past not necessarily because they default more. In another example, a fraud model flags transactions from certain locations more often because historical fraud investigations focused more heavily there, not because fraud is truly higher there.	• Indirect discrimination occurs when the system does not use protected characteristics directly, but relies on proxy variables (e.g., location, school attended, device type, language, or transaction patterns) that are strongly correlated with protected groups, leading to systematic disadvantage even when the model appears "neutral." • Implement controls to detect and reduce proxy-driven harm, including proxy feature review and restriction, subgroup performance testing (including intersectional slices), threshold/abstention rules for high-impact decisions, and documented justification for any high-risk variables used.

Why this matters in sociotechnical systems (not just "the model")

AI sits inside real-world processes, data collection, human decision-making, and unequal access to services. If the surrounding system has gaps (documentation barriers, connectivity issues, biased enforcement, unequal customer support), AI can unintentionally scale those gaps. That is why fairness is not only a technical issue: it is shaped by the whole workflow from data to deployment as well as the ecosystem in which it unfolds.

To help startups check where bias can enter, technical fairness can be broken down into six practical areas: data fairness, application fairness, model design and development fairness, metric-based fairness, system implementation fairness, and ecosystem fairness[28].

Each area includes simple examples so teams can recognize risks early, even before they apply detailed tools and procedures that are covered in the risk assessment section.

28 Leslie (n25)

d.

Data Fairness

Data fairness refers to whether training/validation data is sufficient, reliable, representative, and fit-for-purpose for the populations and contexts where the system will operate.

Startups should treat dataset provenance, representativeness, and label validity as fairness-critical requirements, not just focus on data quality hygiene.

Historical and social bias

Bias or discrimination emerges when historical and social bias becomes "ground truth." If your labels come from decisions that were already unfair (even unintentionally), the model will learn that pattern and repeat it at scale.

For example, if loan approvals in the past favoured formal salaried workers, your dataset may label informal traders as "high risk" simply because they had less access to credit history and not because they are less creditworthy.

Representation gaps

This occurs when the data has representation gaps (who is missing or under-sampled). If some groups are missing or underrepresented, the model would work well for the "majority" in the data and poorly for everyone else.

For example, a customer-service chatbot trained mostly on English queries may fail for Kiswahili-heavy users, rural users, or local language mixed phrasing.

Measurement and label problems

If the data has measurement and label problems (inconsistent definitions, noisy labels, or proxy labels) it can encode structural disadvantages.

For example, using "smartphone type" or "data usage" as a proxy for "ability to pay" can disadvantage low-income customers, even if you never use income as a variable in your model directly.

e.

Application fairness

Application fairness is about whether the AI use case itself is appropriate and justified, given the stakes, the decision context, and who bears the risk (including those indirectly affected). It also includes whether the system's intended purpose is clear and constrained, and whether out-of-scope uses are prevented.

Bias or discrimination can enter when the goal is defined too narrowly (e.g., "reduce cost" only). If you optimize mainly for efficiency, you may create unfair outcomes for certain communities.

For example, a micro-lending SME uses AI to reduce defaults and decides to automatically reject applicants from areas that historically had more late payments. This can become indirect discrimination if those areas correlate with poorer communities or historical underinvestment.

Bias or discrimination can also enter when you try to "automate" a problem that is actually institutional or policy-based. If the real issue is weak processes, poor service coverage, or unequal access, AI can make it worse by adding a "technical stamp" to an existing problem.

For example, using an AI "risk score" to ration health or social support without addressing unequal access to clinics, internet, or documentation.

Organizations should define and document: (i) the decision being supported or automated, (ii) affected stakeholders, (iii) plausible misuse/out-of-scope use, and (iv) fairness expectations before selecting data or models.

f.

Model design and development fairness

Model design and development fairness covers fairness-relevant choices across problem formulation, feature engineering, model selection/training, and testing/validation.

Model features and optimisation can introduce bias and/ or discrimination when you use proxy features that stand in for protected attributes. Even if you remove sensitive attributes, other variables can "act like them."

For example, Location: postcode|estate|ward can correlate with ethnicity, income, or marginalisation patterns, school attended can correlate with class and access, device type can correlate with income and language can correlate with region and community.

Model objective functions and thresholds can also systematically disadvantage one group (e.g., minimising overall error may shift error burden to smaller/under-represented groups).

For example, a fraud model that is optimised to reduce total fraud losses may wrongly block legitimate payments more often for new mobile-money users or smaller merchants.

Lastly, evaluation blind spots, where testing only on aggregate performance can hide subgroup harms. The model can look good on average but fail badly for particular counties, languages, genders, disability groups, or rural users.

Startups should design validation rules to check who the model works for, not only whether it works "on average," and treat proxy risk as a routine design constraint.

g.

Metric-based fairness

Metric-based fairness uses formal statistical definitions (e.g., parity of selection rates, error rates, or predictive values across groups) as a yardstick for assessing discrimination risks.

However, bias and discrimination emerge when metrics mislead performance. For example, different fairness metrics can conflict (e.g., equalised odds can clash with calibration or parity of predictive values), so "passing" one metric does not guarantee fairness in a broader sense.

For example: You can equalise approval rates across groups both privileged or underprivileged but still assign higher interest rates or less favourable loan terms to the group with the higher predicted default risk, thereby satisfying demographic parity at the decision level while imposing unequal economic burdens in practice.

Another view: You can also equalise approval rates across groups and keep loan terms unchanged, but if one group has higher default risk due to structural constraints, this can result in more under-resourced borrowers receiving loans they cannot realistically afford leading to higher delinquency, penalties, and a reinforcing cycle of debt.

Additionally, many metric-based methods require sensitive attribute data and reliable population baselines; those may be unavailable, unreliable, or risky to collect.

Lastly, not focusing on intersectionality may overlook harms concentrated at intersections (e.g., young women in informal employment; rural micro-merchants with limited connectivity) that are obscured when reporting only single-attribute comparisons.

Startups should treat metrics as diagnostic signals (useful but partial), and align metric selection with the real-world decision context and stakeholder expectations.

h.

System implementation fairness

This concerns fairness risks that arise after model development that is during user training, interface design, operational use, monitoring, and feedback loops.

Bias and discrimination emerge through decision-automation bias (overreliance), where users defer to system outputs because they assume the system is objective or superior, reducing critical judgement and increasing the chance that biased outputs translate into discriminatory decisions.

For example, a recruitment tool that flags candidates as "low fit" and HR rejects them without review, even when the model learned biased patterns from past hires.

The opposite may occur where users have automation-distrust bias and ignore legitimate model contributions, producing inconsistent or selectively applied decisions.

For example, one loan officer overrides the model for "known customers," while another doesn't, leading to inconsistent outcomes indirectly discriminating against individuals.

Lastly, predictions based on population correlations, can obscure individual circumstances particularly harmful in high-stakes contexts. A model may use average patterns that do not fit individuals, especially in high-stakes decisions.

For example, a farmer in a remote area is rated "high risk for loan default" due to sparse data coverage, despite having strong repayment behaviour.

Startups should plan for implementer training on model limitations and cognitive biases, and design interfaces that encourage informed judgement (not blind compliance).

Ecosystem fairness

Ecosystem fairness recognises that AI lifecycles are embedded in economic, legal, cultural, and political structures. These structures influence what gets built, who benefits, and whose harms are overlooked.

Innovation agendas and funding patterns can prioritise majority interests and commercial incentives at the expense of marginalised groups, reinforcing structural inequities.

For example, startups may focus on urban users with good data, leaving rural and marginalised communities behind even when the product claims broad inclusion.

On deployment, access barriers (devices, connectivity, language, disability access) can create systematically unequal benefits even when the model is "technically neutral."

For example, digital-only services exclude users with low smartphone access, low data bundles, or limited connectivity.

Startups should treat ecosystem factors as part of product risk management, especially for scaling decisions, target markets, language/localisation, and partnerships that shape who is included or excluded.

Human-Centred Values

Human-centered values serve as the primary philosophical and operational anchor for RAI, ensuring that technology functions as an instrument for human empowerment rather than a replacement for human agency. [29] For African startups, adopting a human-centered approach means prioritizing human dignity, autonomy and rights as the ultimate success criteria for any innovation.[30] This principle requires that AI systems are developed, deployed and monitored in a way that respects the unique moral status of individuals and the collective wellbeing of their communities. By placing the human experience at the core of the innovation lifecycle, startups move beyond building "black box" algorithms toward creating collaborative systems that augment human intelligence while safeguarding the dignity of the people they serve.[31]

From a legal perspective, human-centered values are grounded in international human rights frameworks and constitutional protections for human dignity. This translates into a requirement for startups to ensure their systems do not undermine the rule of law or democratic values throughout the entire product lifecycle. In many African contexts, this legal foundation is enriched by the philosophy of Ubuntu, the principle of "I am because we are", which emphasizes relationality and social cohesion over pure individualism.[32] For a startup Toolkit, this means legal compliance is not just about data privacy, but about ensuring that AI-driven decisions respect the interconnectedness of individuals and do not cause social exclusion or epistemic injustice.[33] This approach aligns AI adoption with African communal values, which often confer a duty for the community to preserve both the social and environmental fabric.[34]

Technically, operationalizing human-centered values requires the implementation of robust oversight mechanisms that allow for meaningful human control.[35] These are typically categorized into three specific interaction modes: **Human-in-the-loop**, which requires active human intervention and approval before a system can execute pivotal decisions; **Human-on-the-loop,** enabling humans to monitor the system in real-time and intervene if the AI deviates from its intended path; and **Human-in-command**, ensuring that humans retain the ultimate authority to decide when and how a system is used.[36]

29 Stefan Schmager, Ilias O Pappas & Polyxeni Vassilakopoulo, Understanding Human-Centred AI : a review of its defining elements and a research agenda [2025] Behaviour & Information Technology , Volume 44, Issue 15

30 Catherine Regis, Jean-Louis Denis, Maria Luciana Axente & Atsuo Kishimoto , Human-Centred AI : A Multidisciplinary Perspective for Policy-Makers, Auditors and Users : *Malwina Anna Wojcik, Toward Addressing Inequality and Social Exclusion by Algorithms : Human-Centric AI through the Lens of Ubuntu* (2024 1st edition Chapman and Hall/CRC)

31 Emmanouil Papagiannidis, Patrick Mikalef & Kieran Conboy, Responsible artificial intelligence governance : A review and research framework [2025] The Journal of Strategic Information Systems, Volume 34, Issue 2

32 Damian Okaibedi Eke, Kutoma Wakunuma & Simisola Akintoye, Responsible AI in Africa: Challenges and Opportunities; *Emma Ruttkamp-Bloem, Epistemic Just and Dynamic AI Ethics in Africa* (2023 Palgrave Macmillan)

33 ibid

34 Jahaziel Osei Mensah & Aimee Van Wynsberghe, Where are the missing values : an exploration of the need to incorporate Ubuntu values into African AI Policy [2025] AI and Ethics

35 Leila Methani, Andrea Aler Tubella, Virginia Dignum & Andreas Theodorou, Let Me Take Over: Variable Autonomy for Meaningful Human Control [2021] Frontiers in Artificial Intelligence , Volume 4

36 ibid

Furthermore, startups must focus on UI/UX design patterns that promote transparency, such as "showing the work" to explain an AI's reasoning or providing discoverable override controls.[37] These technical solutions are essential for users to exercise their agency and make informed decisions about when to interrupt an automated process, thereby maintaining human independence.[38] Centering these values allows African entrepreneurs to ensure cultural and linguistic appropriateness, ensuring that diverse stakeholders can interact with technology on their own terms.

37 Joydeep Chandra & Satyam Kumar Navneet, Advancing Responsible Innovation in Agentic AI : A study of Ethical Frameworks for Household Automation [2025] arXiv (Cornell University)

38 Methani (n40)

Accountability

Accountability serves as the overarching framework for Responsible AI, acting as the essential bridge between abstract ethical values and the operational reality of African innovation.[39] For ventures navigating the complexities of the African market, establishing a clear chain of responsibility is not merely a compliance exercise but a strategic asset that builds the necessary trust with local communities, regulators and international investors who are often wary of the risks associated with emerging technologies in fragmented markets.[40] It requires an organization to be answerable for both the technical performance and the societal impact of its algorithms, ensuring that there are clear consequences and pathways for redress when systems fail or deviate from their intended purpose.[41]

In the foundational stages of development, startups should adopt "Accountability by Design," utilizing participatory processes to align with the relational ethics of Ubuntu, where individual progress is inextricably tied to community flourishing.[42] This involves documenting data provenance and ensuring "meaningful engagement" with stakeholders to prevent systemic biases, such as those identified in historical South African loan datasets that excluded Black women.[43] By adopting this proactive stance, developers prevent the replication of social exclusions and foster the integrity necessary for initial market entry

39 Claudio Novelli, Mariarosaria Taddeo & Luciano Floridi, Accountability in Artificial Intelligence : What It Is and How It Works [2024] AI & SOCIETY Volume 39, issue 4

40 Damilola Oluwaseun Ogundipe and Emmanuel Adeyemi Abaku, Theoretical insights into AI product launch strategies for start-ups : Navigating market challenges [2024] International Journal of Frontiers in Science and Technology Research

41 Novelli (n43)

42 Naadiya Moosajee , Fix AI's racist, sexist bias (14th March 2019) Mail & Guardian <https://mg.co.za/article/2019-03-14-fix-ais-racist-sexist-bias/relevant local knowledge>

43 Damian (n36)

As systems transition toward wider deployment, accountability matures into a robust framework of technical and legal oversight.[44] Increased scale necessitates automated mechanisms, such as comprehensive logging and internal audits, to act as a digital "black box" for reconstructing AI decisions.[45] The Ubuntu perspective recognizes that intelligence includes empathy and "the heart," scaling ventures must maintain meaningful human oversight to protect human dignity.[46] The legal challenges surrounding Kenya's Huduma Namba rollout, which faced significant hurdles for failing to conduct mandatory data protection impact assessments,[47] highlight the risks of bypassing transparency. Ultimately, these tools ensure that growth remains legally resilient and ethically grounded in the collective wellbeing of the community.

44 Novelli (n45)

45 Laura Waltersdorfer, Fajar J, Ekaputra, Tomasz Miksa & Marta Sabou, AuditMAI: Towards An Infrastructure for Continuous Auditing [2024] Computers and Society <arXiv:2406.14243>

46 Jahaziel (n38)

47 Citizenship Rights in Africa Initiative, Governing ID;Kenya's Huduma Namba Programme [2020] <https://digitalid.design/evaluation-framework-case-studies/kenya.html>

Data Privacy and Security

Data privacy and security function as the collective ethical infrastructure of Responsible AI in Africa, serving as a shield for relational dignity and a catalyst for local data sovereignty.[48] In the foundational stages of a pre-seed venture, these principles are often operationalized through the doctrine of meaningful engagement.[49] This approach allows small teams to navigate a significant personnel bottleneck, where specialized technical expertise is scarce by prioritizing deliberative dialogue with communities rather than resource-heavy technical audits. At this level, data governance serves as a set of human-centric protocols that ensure data collection reflects local values, preventing the extractive patterns of data colonialism where information is harvested without returning value to the originating population.[50] This proactive stance is critical for identifying and mitigating context-specific risks, such as the algorithmic bias prevalent in fintech credit scoring platforms.[51] Many AI-powered lending tools rely on alternative data, such as social media activity and mobile metadata, which can create a digital footprint bias that disproportionately

48 Damian Okaibedi Eke, Kutoma Wakunuma, Simisola Akintoye & George Ogoh, Trustworthy AI , African Perspectives : *Seydina Moussa Ndiaye, Building Trustworthiness as a Requirement for AI in Africa: Challenges, Stakeholder and Perspectives* (2025 Palgrave Macmillan)

49 Aishat Oyenike Salami, Artificial intelligence, digital colonialism and the implications for Africa's future development [2024] Data for Policy Proceedings

50 ibid

51 Olajide Babarunde Taofeek, Ekechi Chijioke Cyriacus, Popoola Taoheed Olawale, Adeshina Oguntoye Geroge , Ayittey Selasi & Ogueji Peter Chika Ozo, Machine learning for financial inclusion in agriculture : A study of AI-based credit scoring tools in rural Nigeria [2025] World Journal of Advanced Research and Reviews

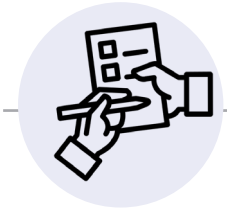
penalizes rural or low-income populations with inconsistent internet access.[52] By embedding governance into the initial design, startups can ensure that their technologies foster genuine financial inclusion rather than entrenching existing socio-economic exclusions.

As a venture transitions into the scaling phase, data governance must evolve from an internal culture into a formal, multi-layered system that accounts for a rapidly shifting regulatory landscape. With over thirty-two African nations having enacted specific data protection laws, scaling requires alignment with both national legislations and broader frameworks like the African Union Data Policy Framework.[53] Maturity in this phase necessitates the performance of Data Protection Impact Assessments to mitigate the risks of unauthorized access and maintain a license to operate across diverse jurisdictions. Technical security measures, such as end-to-end encryption and comprehensive logging, must be integrated into the governance lifecycle to act as a digital black box for auditing decisions and maintaining transparency.

By weaving privacy, security and governance into every stage of growth, African innovators ensure that their technologies are not only legally resilient but also ethically committed to the collective wellbeing of the communities they serve. This integrated approach allows startups to build trust and ensure that the benefits of artificial intelligence are distributed equitably across the continent.

52 Michlle Seng Ah Lee, Luciano Floridi & Jatinder Singh, Formalising trade-offs beyond algorithmic fairness : lessons from ethical philosophy and welfare economics [202] AI and Ethics

53 African Union Data Policy Framework
<https://au.int/sites/default/files/documents/42078-doc-DATA-POLICY-FRAMEWORKS-2024-ENG-V2.pdf>



Transparency and Explainability

For African startups, the principles of transparency and explainability serve as the functional foundation for securing market trust and navigating the continent's intricate regulatory landscape.[54] In the initial stages of technical validation and prototyping, this commitment is best established through the documentation of the data lifecycle, ensuring that the provenance, procurement methods and potential biases of datasets are clearly recorded.[55] This practice is deeply resonant with the value of Ujamaa, which emphasizes a communalist approach to resources and the rejection of extractive data practices.[56] By treating data not as a commodity for exploitation but as a shared asset developed through a relational lens , early-stage ventures can build a culture of data sovereignty where

54 Abaku (n44)

55 Delaram Golpayegani, Isabelle Hupont, Cecilla Panigutti, Harshvardhan J Pandit, Sven Schade, Declan O'Sullivan & Dave Lewis, AI Cards : Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act [2024] Computers and Society <arXiv:2406.18211>

56 Teanna Barrett, Chinasa T Okolo, B Biira, Eman Sherif, Amy Zhang & Leilani Battle, African Data Ethics : A Discursive Framework for Black Decolonial AI[2025] FAccTProceedings of the 2025 ACM Conference on Fairness, Accountability and Transparency

communities remain the primary beneficiaries of the resulting insights. Standardized tools like datasheets or model factsheets provide a proactive mechanism to address algorithmic fairness, establishing credibility with early partners and investors from the outset.[57]

As a solution gains broader traction and encounters diverse institutional voids across borders, the focus naturally shifts toward functional explainability to maintain the intelligibility of outcomes for a wider range of stakeholders. This evolution involves the integration of interpretability assistants that provide localized explanations for specific model predictions.[58] Such technical disclosures must be translated into human-interpretable evidence, allowing founders to build a narrative of accountability that is both technically sound and socially persuasive. This approach ensures that users remain autonomous and are not subject to opaque decision-making processes, which is vital for fostering long-term adoption in sensitive sectors like fintech or healthcare.

Ultimately, these principles are most effective when rooted in a human-centered design philosophy that reflects local relational values. While Ubuntu offers a moral ground of interconnectedness and shared humanness, incorporating Ujamaa further strengthens the ethical framework by prioritizing communal well-being and collective responsibility in the development process.[59] Integrating these philosophies ensures that AI systems are designed to respect the diverse identities and linguistic nuances of the continent.[60] By building ethical reflection into the heart of the product development lifecycle, asking critical questions about social impact and group cohesion during the building phase, startups can ensure their solutions are culturally appropriate and equitable. This integrated approach transforms RAI from a conceptual burden into a tangible strategic advantage, creating a model for innovation that is both sustainable and deeply connected to African social realities.

57 Delaram (n59)

58 Jana Gerlach, Paul Hoppe, Sarah Jagles, Luisa Licker & Micahel H. Bretiner, Decision support for efficient XAI services- A morphological analysis, business model archetypes, and a decision tree [2022] Electronic Markets

59 Teanna (n59)

60 ibid

Contestability

Contestability functions as a foundational pillar for RAI within the African startup ecosystem, providing a critical mechanism for accountability in environments where regulatory frameworks may still be evolving. In the early stages of technical validation and prototyping, this principle is practically applied through “contestability by design,” where developers embed feedback loops and manual override capabilities directly into the system architecture.[61] By establishing these mechanisms during initial development, ventures can build a base of institutional trust, proving to investors and early adopters that their technology is not a “black box” but a responsive tool capable of course correction.[62]

61 Kars Alfrink, Ianus Keller, Geird Kortuem & Neelke Doorn, Contestable AI by Design : Towards a Framework [2022] Minds and Machines

62 ibid

As a venture transitions toward broader market deployment and navigates more complex institutional voids, contestability matures into a strategic risk management and governance tool.[63] In sensitive domains such as fintech and healthcare, the integration of human-in-the-loop oversight ensures that automated decisions do not occur in a vacuum, providing clear pathways for user redress and procedural justice.[64] This focus on human autonomy and the right to challenge AI-driven outcomes allows startups to differentiate themselves from opaque global platforms, building a competitive advantage rooted in reliability and user empowerment. Ultimately, prioritizing contestability ensures that as a startup scales, its technology remains subordinate to human judgment, fostering a model of innovation that is both operationally resilient and ethically robust. This framework allows startups to position contestability as a tool for both technical integrity and market differentiation.

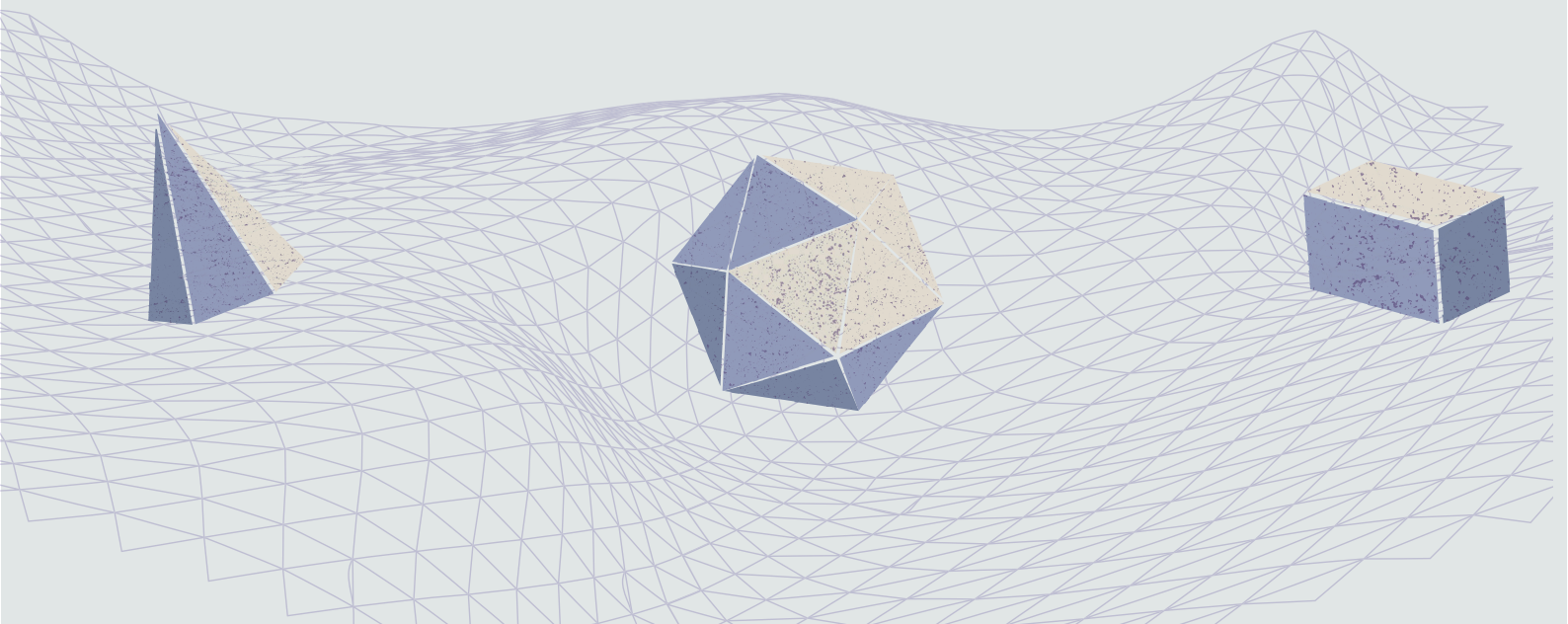
63 Siezte Kai Kuilman, Luciano Cavalcante Siebert, Stefan Buijsman & Catholjin M Jonker, How to gain control and influence algorithms : contesting AI to find relevant reasons [2024] AI and Ethics

64 ibid

02

Module 2

Regulatory Compliance and Governance in African AI Landscape



Objective

AI startups are an increasingly influential force within the global technology sector, making the ethical frameworks they adopt critical to shaping our collective future.[65] Yet, the path to responsible innovation is uniquely challenging for these nascent firms. Existing research highlights a significant gap: while the societal importance of startup ethics is clear, there remains limited understanding of the specific social, cultural and economic pressures that influence their decision-making.[66] Successfully navigating this complex environment is fundamental, as an in-depth grasp of these inter-organizational dynamics must precede the creation of effective, practical ethical guidelines.[67]

To bridge this critical gap, this module moves from analysis to action. We provide a concrete roadmap for integrating robust regulatory compliance covering data protection, intellectual property, and principled ethical governance directly into startup operations. Our objective is to empower innovators with pragmatic strategies for adhering to both Afro-centric values and international standards, thereby transforming regulatory compliance from a perceived obstacle into the very foundation of responsible, trustworthy, and scalable growth.

---

65 Amy A. Winecoff & Elizabeth Anne Watkins, 'Artificial Concepts of Artificial Intelligence: Institutional Compliance and Resistance in AI Startups' [2022] Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society

66 ibid

67 ibid

I. Artificial Intelligence and Data Protection

AI models/ systems may vary in type and application, however a common feature is the presence of data as minimum functioning requirement.[68] At the heart of the AI value chain lies data. For the majority of models, the journey from algorithm to useful application is guided by the training data, which shapes their real-world relevance and utility.[69] However, when data is trained on vast diverse sets of data that originate from various sources such as the web scraping of publicly available text or images, there is controversy around whether personal or sensitive data is included, triggering the application of data protection laws.[70] Personal data is defined as any information that directly or indirectly relates to an identifiable individual.[71] This includes not only direct identifiers like names, but also indirect ones, such as unique traits that can be attributed to a specific person.[72] Bearing this, we assess the applicability of data protection laws throughout the AI lifecycle.

Privacy Integrity in Early Development

Prior to the construction of any AI system or model, identifying its purpose is paramount to ensuring compatibility with the processing of personal data.[73] For instance, an AI system designed for children's interests must inherently be compatible with the processing of children's personal data.[74] This foundational step allows for informed decisions regarding the implications of how startups frame their AI systems' purpose. Following this, establishing the lawfulness of processing personal data is critical, wherein startups must map the applicable legal requirements to establish a valid legal basis for all data processing activities. This is a complex but essential task, guided by various legal instruments ranging from continental to domestic levels within the African setting and beyond.

This section provides crucial guidance on these initial, foundational data protection considerations. It is designed to assist African startups in defining purpose, selecting appropriate legal bases from the very outset of their AI initiatives. These legal bases include consent, contract, legal obligation, vital interests, public tasks or legitimate interests.[75]For every processing activity your startup undertakes, the appropriate legal bases should be identified and documented clearly.

To assist you in this crucial task, the following tables detail key data protection risks inherent in the AI development stage's inception phase, with a focus on the legal bases of consent and legitimate interest, both of which require careful evaluation to ensure compliance with applicable data protection laws. They are designed to be a practical reference, offering insights into specific challenges and how to address them.

68 Andrés Guadamuz, A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs [2024] GRUR International, Journal of European and International IP Law, 2.

69 ibid

70 Jacqueline Poucette, The Right to Be Forgotten and AI ; Legal Obligations, Technical Limits, and Innovation(Lund University 2025, Graduate Thesis) 31

71 Michèle Finck, Frank Pallas, They who must not be identified; distinguishing personal from non-personal data under the GDPR, International Data Privacy Law, Volume 10, Issue 1, February 2020

72 ibid

73 Marco Almada, Law & Compliance in AI Security & Data Protection; AI and Data Protection Training Module , European Data Protection Board [2024] <https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/law-compliance-ai-security-data_en>

74 ibid

75 Lawful bases for data processing include: Consent (unambiguous, informed agreement); Contract (necessity for performance or pre-contractual steps); Legal Obligation (compliance with statutory duties); Vital Interests (protection of a natural person's life); Public Task (performance of official authority or public interest); and Legitimate Interests (balanced organizational needs that do not override individual rights)

Exploring Consent as a Legal Basis for African Startups

In the inception stage of a startup, the development of robust consent mechanisms serves as a foundational element for both legal compliance and ethical AI deployment. By designing granular and tiered consent structures, businesses can satisfy stringent legal requirements for "specific" consent while simultaneously fostering a deeper sense of trust and confidence among their users. This approach empowers individuals through transparent communication and easily accessible withdrawal options, which ultimately builds a more positive and sustainable relationship between the startup and its customer base.

The following table provides a reference point of exploring consent (where applicable) as a legal basis for the processing of data.

Actionable Step for Startups: This column outlines concrete steps that your startup should take to ensure legal compliance during the inception stage, particularly concerning consent.

Key Implementation Details: Here, we expound on the specific details and best practices for executing each actionable step effectively within your AI system design.

Why it's Critical: This section clarifies the importance of each step, explaining its relevance to major data protection laws (e.g., EU GDPR, AU Malabo Convention, Kenya's DPA, Nigeria's NDPA, South Africa's POPIA) and how it safeguards against legal pitfalls and builds user trust.

Actionable Step for Startups	Key Implementation Details	Why it's Critical
1. Implement Granular Consent Controls	Design user interfaces and consent flows to allow your users to selectively agree/disagree to *different* processing operations and purposes. Avoid "all or nothing" prompts.	Essential for "specific" consent and user control. Prevents broad, invalid consent claims where users don't understand what they are agreeing to.
2. Deploy a Layered "Simple Starter" UI to Build Foundational User Trust	Launch with a primary consent choice for core functionality, complemented by a "Learn More" toggle that allows users to opt-in to secondary tiers like AI training or storage.	Establishes transparency early, preventing technical debt and making the brand a "trustworthy user" of data from day one.
3. Ensure Explicit & Unambiguous Opt-in	Consent collection must mandate a clear affirmative act for consent (e.g., ticking an unchecked box, active confirmation). *Never rely on silence, inactivity or pre-ticked boxes.*	Core requirement for valid consent: EU GDPR, South Africa POPIA, AU Malabo Convention.

Actionable Step for Startups	Key Implementation Details	Why it's Critical
4. Build Easy Withdrawal Mechanisms	Ensure your users can withdraw consent at any time, as easily as they gave it. Clearly communicate this right and make the process technically feasible within your system.	Fundamental user right under GDPR. Kenya's DPA allows withdrawal and erasure without delay and Nigeria's NDPA grants this right.
5. Prioritize Clear User Information	Before obtaining consent, provide your data subjects with concise, easy-to-understand information on data types, processing, access, security, retention and their rights.	Critical for "informed" consent. Builds user trust and demonstrates transparency.
6. Design Special Consent for Sensitive Data	For sensitive personal data (e.g., health, biometric and genetic data), ensure your consent mechanisms meet the highest standard of explicit consent, as processing such data generally faces stricter requirements.	EU GDPR defines "special categories of personal data" requiring explicit consent. Nigeria's NDPA, South Africa's POPIA, and Ghana's DPA also make special provisions for sensitive data.
7. Account for Children's Data & Consent	If your service processes children's personal data, implement age verification and obtain consent from a parent or legal guardian. Present information in a child-friendly format.	Mandated in Nigeria and Egypt. General Comment No. 25 emphasizes presenting information in a language children can understand.
8. Plan Consent for Cross-Border Transfers	If your startup anticipates transferring data across national borders, ensure your consent model accounts for additional requirements, which may include specific consent for transfer.	Cross-border data sharing has stricter requirements (e.g., Botswana DPA, South Africa POPIA, Kenya's DPA).

Navigating Legitimate Interest as a Legal Basis for African Startups

Legitimate interests allow for data processing when it is necessary for the purposes of the interests pursued by the controller or a third party, provided these do not override the data subject's rights and freedoms. Applying this basis requires a "balancing test" to ensure the processing is necessary and that the individual's interests are protected. In jurisdictions like Kenya and Nigeria, critics note that while this basis is available, there is a lack of a clearly articulated evaluative framework compared to the GDPR, which can lead to its use in justifying more intrusive data practices.

To apply this basis, controllers must perform a three-part Legitimate Interest Assessment:

Purpose Test: Identify a lawful, real and transparent interest being pursued.

Necessity Test: Confirm the processing is essential and the least intrusive method to achieve the goal.

Balancing Test: Weigh the controller's interests against the data subject/ user's rights, reasonable expectations, potential risks and data sensitivity.

Actionable Step for African Startups: This column provides specific, practical steps for startups considering "legitimate interest" (where applicable), as a legal basis for data processing, detailing the assessments and documentation required.

Key Implementation Details: Here, we offer in-depth guidance on how to perform a comprehensive Legitimate Interest Assessment and address specific challenges related to this basis, especially within African legal contexts.

Why it's Critical: This section outlines the legal imperatives and compliance burden associated with relying on legitimate interests, highlighting why rigorous assessment and documentation are essential, particularly given variations in African legislative clarity on this basis.

Actionable Step for African Startups	Key Implementation Details	Why it's Critical
1. Conduct a Comprehensive 3-Part Legitimate Interest Assessment (LIA)	If relying on "Legitimate Interests," rigorously perform and document the three-part test.	Mandated by various African legislations, this structured approach is critical. You bear the burden of proof that compelling legitimate interests override the data subject's rights.
2. Document All LIA Decisions Thoroughly	Maintain detailed records of your legitimate interest assessments, including rationale, safeguards, and risk mitigation strategies.	Crucial for demonstrating compliance. Nigeria requires documented proof that legitimate interests do not infringe upon fundamental rights or reasonable expectations.
3. Be Wary of Unarticulated Frameworks	Recognize that some African laws may offer less detailed guidance on legitimate interests.	Kenyan DPA and Nigerian NDPA have been noted to lack a clearly articulated evaluative framework for legitimate interest compared to GDPR.
4. Address High Thresholds for AI Profiling	Legitimate interests face a very high threshold. You may need to demonstrate "compelling" interests that *do not* override privacy and potentially notify regulators.	In Botswana, legitimate interests are only valid if "compelling" and not overriding privacy rights, requiring regulator notification for automated processing.
5. Safeguard Automated Decisions	If your AI makes significant decisions (profiling, legal impact), plan for *human oversight, transparency and robust safeguards*.	South Africa's POPIA confers a right not to be subject to solely automated decisions for profiling, allowing automated processing only under certain conditions with human intervention and safeguards. The EU AI Act requires bias detection/correction for sensitive data in high-risk AI.

Data Protection in AI Development Stage

The development of AI is a sequential process where data protection risks are not isolated incidents but rather inherent tensions that accumulate and evolve at each stage. Data protection must be interwoven into every stage of AI development, from initial data sourcing and model design to deployment and ongoing maintenance. This proactive integration, often termed "Privacy-by-Design," prevents costly re-engineering and safeguards against the ethical and legal pitfalls of retroactive remediation. As AI systems become increasingly prevalent across Africa, driven by local priorities and burgeoning tech ecosystems, the need to embed data protection at the core of their operations becomes even more critical.

This section systematically outlines key data protection checks that African startups must appreciate and implement during the development phase. It highlights how core technical issues in AI model training and preparation, such as bias stemming from unrepresentative data or inadequate security measures, directly correlate with legal and compliance requirements across the continent. These include mandates from Nigeria's NDPA, South Africa's POPIA, and Kenya's DPA/KEBS AI Code. By providing a granular view of these intersections, this framework allows startups to grasp the "why" behind each compliance measure, directly linking technical decisions to potential legal penalties, reputational damage and the erosion of user trust. This integrated approach ensures that AI solutions developed by African startups are not only technically robust and innovative but also ethically sound, legally compliant, and positioned for responsible growth.

Data Protection by Design: Risk Mitigation for AI Development in Africa

To assist you in this crucial task, the following table details key data protection risks inherent in the AI development stage. It is designed to be a practical reference, offering insights into specific challenges and how to address them:

- **Risk Category**: This column identifies distinct areas where AI development can pose data protection threats.
- **African Legal & Compliance Requirements (Country/Industry Narrative)**: This section outlines how specific African legal frameworks (e.g., Nigeria's NDPA, South Africa's POPIA, Kenya's DPA/KEBS AI Code) and relevant industry standards address these issues, detailing the statutory requirements and guidance across the continent. It clarifies *why* this risk demands legal and regional attention.
- **Legal & Compliance Implications and Actionable Mitigation Strategies**: This section details the potential consequences of failing to address a particular risk (e.g., fines, turnover-based sanctions, reputational damage) and, crucially, provides concrete, actionable steps and practical solutions that can be implemented to proactively prevent or address each identified risk. It offers guidance on *how* to manage these risks effectively and ensure ongoing compliance.

By utilizing this table, you can systematically identify potential pitfalls, understand your compliance obligations, and implement robust safeguards, thereby ensuring your AI systems are developed ethically, legally, and effectively.

Risk Category	African Legal & Compliance Requirements (Country/Industry Narrative)	Legal & Compliance Implications and Actionable Mitigation Strategies
Bias from Unrepresentative Data	Nigeria's framework requires data to be accurate and reflective of reality, ensuring outputs are not misleading. South Africa emphasizes fairness and just practices in algorithmic decision-making. Kenya's industry code highlights fairness as a societal concern, making bias both a legal and reputational risk.	Biased outputs may be deemed unlawful, eroding trust and damaging reputation. Poor performance leads to costly rework. **Recommendation:** Startups should curate diverse datasets, conduct bias audits before training, and document dataset limitations and mitigation steps.
Excessive Data Collection & Unclear Retention	South Africa enforces minimization and storage limitation, requiring organizations to collect only what is necessary and retain it for defined purposes. Kenya's framework and industry code demand clear purpose specification and documentation of minimum necessary data, embedding minimization into governance.	Collecting unnecessary data or retaining it indefinitely breaches minimization and storage rules, leading to fines and higher breach risks. **Recommendation:** Startups must define scope clearly, collect only essential data, and enforce automated retention policies with anonymization or deletion once the purpose is fulfilled.
Lack of Proactive Risk Assessment	Kenya's law requires privacy to be embedded into system design from the outset. The Kenyan industry code recommends privacy impact assessments as a proactive governance tool, especially for novel AI technologies.	Absence of proactive risk assessment or DPIAs results in compliance failures, undetected vulnerabilities and expensive retrofitting. **Recommendation:** Instead of starting with a full Data Protection Impact Assessment, perform a lightweight "threshold analysis." This screening determines if a comprehensive assessment is legally required based on your specific data activities.

Risk Category	African Legal & Compliance Requirements (Country/Industry Narrative)	Legal & Compliance Implications and Actionable Mitigation Strategies
Inadequate Security	Kenya's framework demands layered safeguards including technical, operational, and physical protections. Nigeria mandates specific technical controls such as encryption, firewalls and access restrictions. Egypt underscores the severity of failures by attaching criminal liability and fines under its cybercrime regime.	Weak safeguards expose systems to breaches, leaks and theft of intellectual property. Penalties include fines, turnover-based sanction and imprisonment. **Recommendation:** Startups must encrypt data end-to-end, enforce strict role-based access controls, apply differential privacy for sensitive datasets and adopt robust cybersecurity frameworks including firewalls and penetration testing.
Misclassified Anonymity	South Africa treats pseudonymized data as personal if it can be linked back to the data subject, requiring continued safeguards. Kenya defines anonymisation as irreversible, setting a high bar for compliance and ensuring that data subjects cannot be re-identified.	Treating pseudonymized data[76] as anonymous leads to unlawful processing, compliance breaches and erosion of trust. **Recommendation:** Startups should regularly test datasets for re-identifiability, document anonymization techniques, maintain safeguards until data is irreversibly anonymous, and establish clear governance distinguishing anonymous, pseudonymized, and identifiable data.

76 Pseudonymized data is information that has been processed so that it can no longer be attributed to a specific person without the use of "additional information," such as a decoding key

Responsible AI Deployment

The deployment of AI marks the critical transition from design and testing to real-world application; a phase where theoretical data protection safeguards meet practical, often unpredictable, operational environments. This section provides the necessary conceptual and regulatory groundwork for understanding and managing data protection risks during AI deployment, particularly within African regimes. It recognizes that Africa's legal landscape is neither uniform nor static; rather, it is characterized by a dynamic patchwork of national laws, regional protocols and pan-continental initiatives, all set against a backdrop of diverse cultural values and uneven enforcement capacities. Effectively navigating this complex environment is essential for compliant and trustworthy AI operations.

To assist you in this crucial task, the following table details key data protection risks. It is designed to be a practical reference, offering insights into specific challenges and how to address them:

Risk Category: This column identifies distinct areas where AI deployment can pose data protection threats.

Core Issue: Here, we expound on the fundamental problem or challenge associated with each risk category, explaining *what* specific harm or breach could occur.

Legal & Compliance Implications: This section outlines how specific African legal frameworks (e.g., South Africa, Kenya, Nigeria) address these issues, detailing the statutory requirements and potential consequences of non-compliance. It clarifies *why* this risk demands legal attention.

Mitigation Strategies: Most importantly, this column provides actionable steps and practical solutions that can be implemented to proactively prevent or address each identified risk. It offers guidance on *how* to manage these risks effectively and ensure ongoing compliance.

By utilizing this table, you can systematically identify potential pitfalls, understand your compliance obligations and implement robust safeguards, thereby ensuring your AI systems are deployed ethically, legally, and effectively within the AI.

Risk Category	Core Issue, Legal & Compliance Implications and Mitigation Strategies
Opaque Automated Decision-Making : AI systems making high stake decisions without transparency erode trust and block statutory rights to challenge outcomes.	High-stakes examples include loan approvals, hiring rejections, healthcare triage, insurance claims, predictive policing, welfare eligibility and bail/sentencing risk assessments. **Legal implication:** South Africa and Kenya require data subjects to be informed about the logic of automated processes and to make representations. **Mitigation:** Use Explainable AI, provide plain-language reasoning and embed human-in-the-loop review for significant decisions.

Risk Category	Core Issue, Legal & Compliance Implications and Mitigation Strategies
Proliferation of AI Hallucinations Factually incorrect outputs derived from personal data cause harm, operational errors and breach data quality principles.	**Legal implication:** Nigeria and South Africa mandate personal data to be accurate, complete, not misleading and kept up to date. **Mitigation:** Deploy Retrieval-Augmented Generation (RAG), establish real-time accuracy monitoring and implement user-facing correction mechanisms. Additionally, a startup may deter from using personal data, where possible, to reduce compliance cost.
Unauthorized Purpose Drift in Live Systems: AI systems expanding functionality beyond approved scope invalidate the original legal basis for processing and breach purpose limitation.	**Legal implication:** Kenya requires explicit, specified, legitimate purposes; Nigeria enforces purpose limitation. **Mitigation:** Maintain dynamic purpose registries, enforce strict change-control protocols and implement purpose-scoped access controls.
Systemic Failure to Execute Data Subject Rights : Inability to comply with erasure or rectification requests undermines statutory rights and triggers penalties.	**Legal implication:** Nigeria and Kenya both enshrine the right to erasure and deletion "without delay." **Mitigation:** Build robust data lineage tracking, invest in machine unlearning and automate workflows for rights requests.
Critical Deficit in Operational AI Competence and Literacy: Lack of skilled personnel and literacy undermines governance, monitoring, and compliance, collapsing safeguards.	**Legal implication:** Kenya's Data Protection Commissioner oversees enforcement, requiring organizational competence. **AI Strategy provisions:** The African Union's AI Strategy calls for states to *"incorporate workforce development and technical skill-building provisions within their respective AI and data governance strategies"* to ensure sustainable competence. The Kenya AI Strategy also addresses *"the critical need for equitable access to AI through developing AI skills across all levels of society"* to ensure fostering of local talent and promotion of AI literacy across all demographics. **Mitigation:** Conduct competency gap audits, Mandate ongoing training or consult in AI ethics and data protection law specialists to create baseline literacy on establishing clear accountability lines for AI.

II. AI and Intellectual Property Rights

The development of Artificial Intelligence presents a critical nexus of legal and technical challenges for African startups, particularly in establishing responsible AI practices. This introductory section will explore the intricate relationship between AI and Intellectual Property within the African context, focusing on key thematic areas crucial for navigating this complex landscape.

First, we will delve into the challenges posed by Text and Data Mining practices and the associated risks of copyright infringement. As AI models learn by identifying patterns rather than direct duplication, profound ambiguities arise regarding the "lawful basis" for processing vast datasets. In many African jurisdictions, including countries like Egypt, the absence of explicit statutory "fair use" or "fair dealing" safe harbors for commercial TDM creates a precarious environment for AI developers, potentially leading to liability for copyright infringement when AI outputs mirror protected training data. This section will outline the IP risks inherent in TDM and discuss various mitigation strategies, including the importance of bespoke "AI-Ready" licensing, engaging with Collective Management Organizations, adhering to open-license hygiene and conducting public domain verification.

Second, a unique concern within Africa is the potential for cultural misappropriation through the ingestion of Traditional Knowledge systems into AI. The AU- AI Continental Strategy underscores the need to safeguard indigenous knowledge, advocating for the creation of technologies deeply rooted in African values, norms and cultural identities and leveraging TK as a springboard of authentic African innovation and creativity. This inherently seeks to prevent the misuse or unacknowledged extraction of cultural heritage, thereby addressing the underlying concerns of cultural extraction.

Finally, the discussion will extend to the broader IP protection of AI software components. This includes exploring how existing IP regimes, copyright for source code, patents for novel algorithms, and trade secrets for proprietary models and training data, apply to the constituent elements of AI systems within African legal frameworks. Copyright typically protects the expression of source code, while the patentability of AI algorithms often depends on demonstrating a "further technical effect" beyond the abstract algorithm itself .Trade secrets are particularly crucial for protecting proprietary algorithms, model architectures (parameters and weights), and training datasets, as they maintain confidentiality and offer a competitive edge without public disclosure. Understanding these intersections and thematic areas is fundamental for African startups to innovate responsibly and sustainably in the rapidly evolving AI ecosystem.

Intellectual Property Risks in Text and Data Mining (TDM) for AI Development in Africa

To assist you in this crucial task, the following table details key approaches to TDM. It is designed to be a practical reference, offering insights into specific challenges and how to address them:

Identify the approach and scope: Review each row to understand how different jurisdictions frame TDM, whether through fair dealing, research-only exceptions or broader opt-out models. This helps you see where your project fits.

Assess risks and obligations: Use the "Key Actions & Risks" column to recognize potential legal uncertainties, infringement risks or cultural misappropriation concerns. Match these with the relevant national or regional laws noted in the table.

Apply safeguards and strategies: Translate the risks into operational steps: such as securing licenses, conducting dataset audits, establishing prior informed consent with communities or honoring opt-out mechanisms. Document these actions to demonstrate compliance and ethical engagement.

This structure makes the table a decision-making tool: it guides startups and researchers through the complexities of TDM in Africa, helping them align innovation with legal compliance and cultural respect.

Approach to TDM	Key Narratives / Scope	Key Actions & Risks
No Explicit TDM Exception	Copyright frameworks in Kenya and South Africa often lag technological advancements, leading to "fair dealing" provisions that are too narrow for modern TDM. This creates a challenging environment for integrating new technologies like AI.	For both Kenya and South Africa, the primary risk is significant legal uncertainty and high potential for copyright infringement in TDM activities, especially commercial ones. This implies a general need for explicit licenses. South Africa's *Copyright Amendment Bill of 2017* aims to address this by including provisions for AI-created works and expanding fair use.
Fair Dealing with 4-Factor Test	This approach, seen in countries like Nigeria, allows TDM to be evaluated through interpretative principles such as a 4-factor test, considering "transformative use" and market impact. It typically includes exceptions for private or non-commercial research, seeking a balance between innovation and creator protection.	This carries a medium risk, requiring TDM users in countries like Nigeria to audit datasets carefully for compliance with the 4-factor test. There is a recognized liability for extensive ingestion of copyrighted works without proper licensing, necessitating caution for commercial TDM.

Approach to TDM	Key Narratives / Scope	Key Actions & Risks
Research-Only Fair Dealing	This legislative choice permits TDM exclusively for non-commercial research, creating a clear distinction that fosters academic exploration while explicitly excluding commercial TDM. This is evident in countries such as Egypt, which has limited exceptions in its copyright law.	This approach entails a high risk of copyright infringement for any commercial TDM in countries like Egypt. Additionally, data privacy laws introduce another layer of risk, requiring explicit consent for personal data processing under Egypt's data protection framework.
Data Sovereignty & Traditional Knowledge (TK) Protection	Within the broader African context, there's a strong emphasis on leveraging data for intra-African sharing while rigorously safeguarding Traditional Knowledge under IP and cultural rights frameworks. Initiatives like the AU Digital Strategy for Africa and the AU - AI Continental Strategy for Africa drive this, promoting responsible data utilization and localizing personal data on the continent. Traditional Knowledge is viewed as a springboard for authentic African innovation, necessitating a deliberate indigenous knowledge-sensitive continental strategy.	The risk shifts from copyright infringement to potential cultural misappropriation. TDM practices must align with cultural IP goals and mandate consultation with TK holders to ensure ethical engagement. Discussions around the African Continental Free Trade Area's Intellectual Property Protocol highlight championing national interests, particularly in protecting TK and genetic resources. There is also a recognized need for strong IP protections to ensure African researchers are not just data providers but co-creators with legal rights to their data.
Opt-Out for Commercial Use	Modern legislation can permit commercial TDM by default unless copyright holders explicitly opt-out using machine-readable means (e.g., robots.txt). This progressive stance, exemplified by the EU DSM Directive Article 4, facilitates TDM as a standard practice for innovation.	This presents a medium risk. TDM users operating under such frameworks must systematically honor machine-readable opt-outs. Emerging regulations like the EU AI Act also introduce compliance requirements, such as publishing a detailed summary of the content used for training general-purpose AI models, adding layers of transparency and responsibility.

Mitigation Strategies

To assist startups in navigating the complex landscape of Text and Data Mining (TDM) and AI compliance, the following table details key actionable steps. It is designed as a practical reference, offering insights into specific challenges and how to address them effectively:

Actionable Step: This column identifies distinct measures that startups should adopt to ensure lawful and ethical AI deployment.

Legislative Narrative / Rationale: Here, we expand on the statutory and regulatory context underpinning each step, drawing from African legal frameworks such as Nigeria's NDPA 2023, Kenya's DPA and South Africa's POPIA. It explains why each measure is legally significant and what risks it mitigates.

Practical Moves for Startups: Most importantly, this column provides concrete, operational actions that startups can implement immediately. These steps translate legislative requirements into day-to-day practices, ensuring compliance and reducing exposure to infringement or regulatory penalties.

By utilizing this table, startups can systematically identify potential pitfalls, understand their compliance obligations, and implement robust safeguards. This ensures that AI systems are deployed ethically, legally, and effectively within African jurisdictions, while also aligning with emerging global standards such as the EU AI Act.

Actionable Step	Legislative Narrative / Compliance Rationale	Practical Moves for Startups
Direct "AI-Ready" Licensing	Negotiating bespoke licenses establishes a lawful basis for data processing under the NDPA 2023. In Kenya, where statutory clarity for commercial TDM is limited, explicit licensing helps avoid infringement risks.	• Draft contracts with explicit TDM clauses • Align with NDPA 2023 lawful basis • Monitor evolving statutory clarity in Kenya.
Engage Collective Management Organizations (CMOs)	Collective Management Organizations, such as the Music Copyright Society of Kenya and the Copyright Society of Nigeria, are recognized under national copyright laws to administer rights. Engaging them ensures compliance with remuneration obligations and reduces infringement claims.	• Partner with CMOs • Secure blanket licenses for creative works • Document remuneration pathways.

Actionable Step	Legislative Narrative / Compliance Rationale	Practical Moves for Startups
Implement Robust Open License Hygiene	Creative Commons Attribution (CC BY) and Creative Commons Zero (CC0) licenses permit broad reuse. However, Non-Commercial (NC) and No-Derivatives (ND) licenses restrict commercial exploitation. Using NC-tagged data commercially would breach copyright and contract law.	• Automate license filtering • Exclude NC/ND datasets • Maintain compliance logs.
Conduct Regular Public Domain Verification	Copyright terms in many African jurisdictions last for the life of the author plus 50 years. Works beyond this term fall into the public domain. However, data protection laws such as South Africa's POPIA and Kenya's DPA require removal of personal identifiers before use.	• Audit datasets for expired terms • Clean identifiers • Align with South Africa's POPIA and Kenya DPA requirements.
Maintain Comprehensive Copyright Provenance Logs	Metadata records provide evidence of lawful access and licensing. Under frameworks like Nigeria's NDPA and South Africa's POPIA, maintaining provenance supports accountability and auditability.	• Maintain dataset provenance logs • Store license metadata • Use logs for regulatory defense.
Publish High-Level Public Training Summaries	Transparency obligations under Article 53 of the EU AI Act require disclosure of training datasets. Nigeria's NDPC also mandates documented impact assessments. Publishing summaries demonstrates compliance and builds trust.	• Publish dataset summaries • Document impact assessments • Align with EU AI Act benchmarks.
Recognize Opt-Out Mechanisms	Robots.txt and machine-readable tags are recognized mechanisms for rights holders to restrict TDM. Honoring these signals supports fair dealing defenses and prevents unauthorized use.	• Detect opt-out signals • Exclude flagged datasets • Document opt-out compliance
Deploy Anti-Regurgitation Filters	Copyright law prohibits verbatim reproduction of protected works. Similarity checkers prevent AI systems from outputting near-verbatim content, reducing infringement risks.	• Run similarity checks • Block verbatim outputs • Modify flagged responses.

Actionable Step	Legislative Narrative / Compliance Rationale	Practical Moves for Startups
Utilize Human-in-the-Loop for IP Protection	In Nigeria and Kenya, demonstrable human creative input is increasingly required for AI outputs to qualify for copyright protection. This distinguishes transformative works from derivative reproductions.	• Add human review layers • Ensure transformative edits • Document human input.
Implement Watermarking and Attribution	Digital watermarking standards such as the Coalition for Content Provenance and Authenticity (C2PA) help distinguish AI-generated works. This prevents deceptive practices and supports attribution obligations under consumer protection and copyright law.	• Apply digital watermark • Attribute AI outputs • Prevent deceptive practices.

Traditional Knowledge and Cultural Accountability in AI Development

This section provides a detailed guide for African AI startups on navigating the complexities of Traditional Knowledge and ensuring cultural accountability.

Protecting non-formalized intellectual property, such as traditional knowledge and cultural expressions, is a strategic priority underscored by the AU - AI Continental Strategy (2024–2030). This strategy aims to foster responsible AI development deeply rooted in African values, norms, and cultural identities, a framework crucial for guiding AI innovation in a manner that respects and integrates Africa's rich cultural heritage.

To assist you in this crucial task, the following table details key risks related to traditional knowledge in AI. It is designed to be a practical reference, offering insights into specific challenges and how to address them:

Identify and assess risks: Look at each risk category to see where your AI project may face challenges, such as community consultation or benefit-sharing.

Match risks to obligations: Use the legal and compliance column to understand which African laws and protocols apply and what requirements must be met.

Implement safeguards: Apply the mitigation strategies as operational steps, document them thoroughly, and embed ethical principles like Ubuntu to ensure compliance and community trust

The table is a hands-on compliance Toolkit that startups can use to systematically manage risks and align with African centric values and principles.

Risk Category	Core Issue	Legal & Compliance Implications	Mitigation Strategies
Traditional Knowledge Consultation	AI systems may use indigenous data, cultural expressions, or folklore without community consent, leading to cultural misappropriation and loss of trust.	National laws such as Kenya's Protection of Traditional Knowledge and Cultural Expressions Act and regional instruments like the Swakopmund Protocol require Prior Informed Consent before documenting or using TK. Non-compliance can result in legal disputes and reputational damage.	• Establish Prior Informed Consent processes. • Conduct transparent consultations with communities. • Document agreements and ensure communities understand potential uses and benefits. Align with statutory requirements to avoid legal challenges.
Benefit-Sharing Agreements	AI models that generate commercial value from TK risk exploiting communities if benefits are not shared fairly. This can lead to accusations of biopiracy.	Laws such as South Africa's Indigenous Knowledge Act and Kenya's TK and Cultural Expressions Act mandate benefit-sharing when TK is used commercially. The AU - AI Continental Strategy also emphasizes fairness and equity in AI development.	• Draft equitable benefit-sharing agreements. • Provide monetary compensation, community investments or capacity-building initiatives. • Ensure agreements are transparent and enforceable. • Embed ethical principles such as Ubuntu to reinforce trust.

IP Protections: Software in AI systems

The intellectual property landscape surrounding Artificial Intelligence extends significantly beyond the training data to encompass the AI software, algorithms, and models themselves. Protecting these core components is crucial for AI developers and businesses due to their commercial value and role in innovation. The application of traditional IP laws to AI presents several challenges, as the rapid advancement of AI technologies often outpaces the development of IP laws. For instance, while software is generally protected by copyright, some traditional exclusions for computer programs in patent law create legal uncertainty for AI-based products.

Below is a legal mapping and actionable recommendations, with a focus on the African legal context where the provided sources allow, for protecting various components of AI systems:

Identify the component: Use the first column to pinpoint which AI asset you are dealing with (source code, algorithms, models or training data).

Check protections and risks: Review the second and third columns to see which intellectual property rights apply and how they function under African legislative frameworks.

Apply safeguards: Follow the fourth column as a checklist of actions; register copyrights, file patents, enforce trade secret protections, and comply with data protection laws.

This condensed approach makes the table a practical compliance tool, helping startups quickly connect each AI component to its legal protections and required safeguards.

AI Component	Key IP Protection(s)	What It Is & How It Applies to AI	What Your Startup Should Do
Source Code	Copyright	Copyright protects the specific written code, like how text in a book is protected. It covers the exact expression of algorithms, not the underlying ideas or functions. Human creative effort is required, so ownership of AI-generated code may be unclear if human input is minimal. In Africa, frameworks such as South Africa's Copyright Act and Kenya's Copyright Act govern protection of software code.	• Add clear copyright notices in your code. • Register copyright where possible to strengthen legal standing. • Ensure sufficient human input when AI assists in coding. • Follow all open-source license terms carefully.

AI Component	Key IP Protection(s)	What It Is & How It Applies to AI	What Your Startup Should Do
Algorithms	Patents, Trade Secrets	Patents can protect an AI invention if it is new, non-obvious and solves a real-world problem with a technical effect. Abstract mathematical algorithms alone are usually not patentable. Trade secrets protect valuable, confidential algorithms indefinitely if secrecy is maintained and they provide a business advantage. Relevant frameworks include Kenya's Industrial Property Act and South Africa's Patents Act.	• For patents: focus applications on technical uses of algorithms that solve real-world problems. • For trade secrets: enforce strict confidentiality through access controls, NDAs, and encryption. • Document measures to demonstrate compliance with patent and trade secret requirements.
AI Models	Trade Secrets	The design, parameters, weights and training methods of AI models are valuable assets that define capabilities and competitive advantage. Protecting these details is critical. Trade secret protection is recognized under frameworks such as Nigeria's Trade Secrets provisions and South Africa's common law protection of confidential information.	• Maintain strong trade secret protections. • Use strict access controls and NDAs. • Document all confidentiality measures to ensure enforceability. • Regularly review compliance with national trade secret laws.
Training Data	Trade Secrets, Copyright, Database Rights	Proprietary datasets can be trade secrets if kept confidential and commercially valuable. Copyright may apply to data containing creative works or to how data is organized. Database rights may protect structured collections of data. African data protection laws, such as Kenya's DPA and South Africa POPIA, require a lawful basis for processing of personal data. Cross-border data sharing often has additional requirements.	• Establish clear data governance policies. • Use confidentiality agreements and access controls. • Ensure lawful basis for data processing under national data protection laws. • Verify data sources to avoid IP infringement. • Follow database rights and copyright rules when organizing or annotating datasets. • Comply with cross-border data transfer requirements.

III. Algorithmic Bias and Fairness in Africa: Analysis, Law, and Implementation for Startups

The AI landscape in Africa presents a critical challenge in algorithmic bias, demanding that fairness be a central tenet from design to deployment. As AI systems learn from existing data, they risk inheriting and amplifying historical and societal inequalities, with profound implications given Africa's diverse contexts. Key elements of fairness are undermined, for example, through data bias in financial inclusion where loan algorithms, relying on traditional data, have shown gender bias, disproportionately impacting women in informal economies.[77] Similarly, the lack of representational data in facial recognition systems leads to significantly lower accuracy for diverse African populations, posing serious threats of misidentification and exclusion.[78] Furthermore, the absence of linguistic diversity in training data can result in hate speech detection algorithms misclassifying African languages, thereby marginalizing voices and limiting access to information.[79] For African startups, addressing these ingrained biases, from data collection's sampling bias to real-world deployment bias, is not merely an ethical obligation but a strategic necessity. Success hinges on proactively integrating transparency, accountability, and inclusivity throughout the AI lifecycle, ensuring equitable outcomes and building trust across diverse communities.

The following table follows a simplified structured design that connects each AI fairness component to its risks, legal context, and safeguards.

- **Identify the Component**: Pinpoint which fairness aspect of AI you are addressing (e.g., data collection, bias auditing, transparency).
- **Check Protections & Risks**: Review the risks of bias and unfairness, and how they manifest in AI systems.
- **African Legal & Compliance Requirements**: Map each fairness consideration to African legal frameworks (e.g., POPIA, NDPA, Kenya's DPA/KEBS AI Code, AU-AI Continental Strategy).
- **Apply Safeguards**: Use the final column as a checklist of actionable steps to embed fairness and compliance.

77 Notice Pasipamire & Abton Muroyiwa, Navigating algorithm bias in AI: ensuring fairness and trust in Africa [2024] Front.Res. Metr, Anal, Vol 9

78 ibid

79 Aishat Oyenike Salami, Artificial intelligence, digital colonialism and the implications for Africa's future development [2024] Data for Policy Proceedings; AI, Ethics and Policy Governance in Africa

AI Fairness Component	Protections & Risks	African Legal & Compliance Narratives	Safeguards / Actionable Checklist
Data Collection & Representation	Biased or incomplete datasets risk excluding communities, producing unfair outcomes and reinforcing inequality.	In South Africa, the data protection framework requires lawful and fair collection, with strict limits on unnecessary data. Nigeria emphasizes proportionality and necessity, while Kenya's Data Protection Act and KEBS AI Code of Practice highlight transparency and ethical sourcing. Across the continent, inclusivity is a guiding principle, ensuring datasets reflect diverse populations.	• Collect datasets that reflect diverse African populations. • Avoid reliance on foreign datasets that ignore local realities. • Partner with local institutions for representative data. • Document sources, consent, and limitations clearly.
Bias Detection & Auditing	Without systematic checks, AI systems may produce discriminatory results, undermining trust and legitimacy.	South Africa's framework prohibits processing that results in unfair discrimination. Nigeria's approach requires fairness in automated decision-making, while Kenya's KEBS AI Code mandates bias testing and fairness audits. Regional strategies emphasize that fairness must be demonstrable, not assumed.	• Conduct bias audits during training and testing. • Apply fairness metrics (equal opportunity, disparate impact). • Compare outputs across demographic groups. • Engage independent reviewers or ethics boards.
Inclusive Design & Development	Systems risk excluding cultural or linguistic groups if design is not inclusive, limiting accessibility and fairness.	African Union strategies promote people-centric AI, embedding values such as Ubuntu, which highlight interconnectedness and inclusivity. Kenya's KEBS AI Code encourages culturally sensitive design and equitable solutions. These narratives stress that AI must reflect local realities and languages.	• Build diverse, multidisciplinary teams. • Incorporate local languages and cultural contexts. • Test prototypes with varied user groups. • Establish feedback loops for continuous improvement.

AI Fairness Component	Protections & Risks	African Legal & Compliance Narratives	Safeguards / Actionable Checklist
Transparency & Explainability	Opaque decision-making erodes trust and denies individuals the ability to challenge or understand outcomes.	South Africa's framework grants individuals rights to access and correct their data. Nigeria requires disclosure when automated decision-making is used. Kenya's KEBS AI Code calls for transparency and explainability, ensuring AI systems are understandable to non-experts.	• Document training data and decision logic. • Use explainable AI tools. • Provide user-friendly explanations for automated decisions. • Publish fairness and accountability reports.
Accountability & Governance	Weak oversight risks unfair outcomes going unaddressed, leaving users without remedies.	In South Africa, organizations are held responsible for compliance and fairness. Nigeria's framework places obligations on controllers and processors, while Kenya's Data Protection Act and KEBS AI Code mandate accountability and ethical oversight. The AU-AI Continental Strategy emphasizes governance and human oversight.	• Develop internal fairness policies aligned with governance principles. • Assign responsibility for bias monitoring. • Create grievance mechanisms for users. • Align governance with African values of fairness and Ubuntu.
Community Engagement & Feedback	Excluding communities risks cultural misappropriation, loss of trust, and accusations of exploitation.	Kenya's Traditional Knowledge Act requires community consultation and prior informed consent for use of cultural heritage. The Swakopmund Protocol protects traditional knowledge and folklore, while the AU AI Strategy emphasizes Ubuntu values and benefit-sharing. These narratives highlight that communities must be active participants in AI development.	• Consult affected communities before deployment. • Respect traditional knowledge and cultural values. • Provide accessible feedback channels. • Share AI benefits with communities to build trust and legitimacy.

IV. Data Security and Liability Breach

The digital infrastructure across Africa, coupled with AI's inherent vulnerabilities like data poisoning and adversarial attacks, creates a complex threat landscape further complicated by deepfake-driven misinformation, data colonial dependency and compute scarcity. This environment is exacerbated by high corruption levels and weak data infrastructure in some regions, increasing susceptibility to breaches .

Concurrently, a rapidly maturing regulatory landscape features national data protection laws, the AU Malabo Convention, and widely accepted global standards flowing from the influential EU GDPR and the EU AI Act which has extra-territorial applications to AI systems with functions, outputs or impacts affecting the EU jurisdiction. A critical emerging challenge in Africa is the absence of clear legal liability rules for AI harm, often relying on existing consumer protection regimes for recourse. This presents African startups with a multi-layered compliance burden, where a single security incident can trigger liabilities across diverse legal frameworks, compounded by varied technical resources and data limitations across nations .

Therefore, a purely reactive, compliance-focused approach is insufficient. Successful startups must adopt a proactive, governance-led strategy integrating cybersecurity and data protection, grounded in Africa's unique context. This involves ethical data stewardship and culturally sensitive AI governance, like the Ubuntu philosophy, to align AI development with local values. Such an approach fosters operational resilience, builds trust with users and investors, and transforms security into a sustainable competitive advantage.

Below is a structured approach to leveraging the table, focusing on practical application within the African regulatory context:

- **Identify the Data Security Issue**: Utilize the first column ("Key Data Security Issue") to pinpoint the specific challenge your AI startup may encounter, ranging from vulnerable infrastructure and regulatory compliance to the complexities of AI liability.
- **Implement Actionable Safeguards**: Refer to the second column ("Actionable Steps for Startups") as a direct checklist. These provide concrete, practical measures your team can adopt to ensure compliance, effectively mitigate risks and foster trust within your operations and with stakeholders.

This structured approach transforms the table into a highly practical compliance tool, enabling startups to swiftly connect each identified data security issue with its foundational rationale and requisite protective measures.

Key Data Security Issue	Actionable Steps for Startups
Vulnerable Infrastructure & AI-Specific Threats African startups often operate on developing digital infrastructure, which can be more vulnerable to cyberattacks. AI introduces unique risks such as data poisoning, adversarial manipulation and misuse of deepfakes. Legal narratives emphasize that organizations must exercise due diligence in securing systems, while ethical narratives highlight data sovereignty and the need to reduce dependence on external models.	• Deploy strong cybersecurity measures (encryption, access controls, intrusion detection). • Develop and regularly update strategies to counter AI-specific threats. • Invest in secure, localized infrastructure to strengthen sovereignty. • Build local AI capabilities to reduce reliance on external models. • Conduct regular audits and penetration testing.

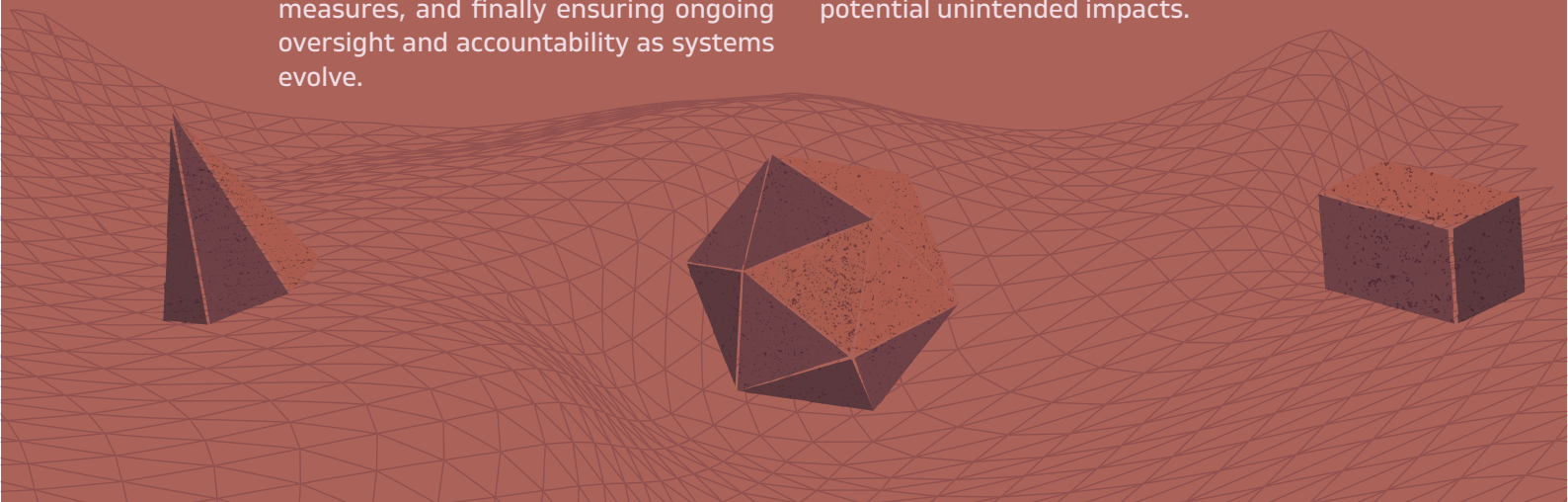
Key Data Security Issue	Actionable Steps for Startups
Regulatory Compliance & Cross-Border Data Flows Startups face a patchwork of national data protection and cybercrime laws, alongside continental initiatives and international standards. Non-compliance, especially with data localization and transfer rules, creates significant risks. Compliance is a legal obligation, not optional, and violations can lead to penalties and reputational harm.	• Establish clear legal bases for all personal data processing. • Ensure adherence to national data protection and cybercrime laws. • Understand implications of international regulations (e.g., GDPR, AI Act) if operating globally. • Conduct regular compliance audits to track legislative changes. • Develop internal policies and training programs for compliance.
Ambiguous AI Liability Many jurisdictions lack clear rules on liability for AI-induced harm. Existing consumer protection laws often serve as fallback mechanisms, but this creates uncertainty. Startups face unpredictable risks if responsibility for AI failures or breaches is unclear, making documentation and proactive risk management essential.	• Conduct thorough risk assessments for all AI systems. • Maintain detailed documentation of AI design, testing and deployment. • Develop clear incident response protocols and accessible redress mechanisms
Enforcement & Penalties for Data Security Breaches Data protection and cybercrime laws are actively enforced across Africa, with severe penalties for breaches. Legal narratives stress accountability, mandatory breach notifications and security measures. Consequences of non-compliance include fines, criminal charges and reputational damage, underscoring the importance of proactive compliance.	• Implement robust security measures across systems. • Develop a comprehensive incident response plan. • Ensure timely breach notifications to authorities and affected individuals. • Conduct regular compliance audits to identify risks early. • Train employees on data security best practices and legal obligations.
Ethical AI Governance & Data Stewardship Ethical guidelines often precede legal mandates but are vital for building trust and mitigating risks. Narratives rooted in African values, such as Ubuntu, emphasize collective well-being, cultural sensitivity and responsible stewardship of data. Ethical governance reduces insecurity and bias in data handling, while strengthening public trust.	• Embed ethical governance frameworks into AI development. • Adopt culturally sensitive design principles, engaging local communities. • Promote transparency in how AI systems use and protect data. • Conduct ethical impact assessments to identify risks proactively.

03
Module 3

Technical AI Risk Assessment and Mitigation Framework

This module presents a framework to support startups in identifying, assessing, and managing risks associated with the development and deployment of AI systems. Rather than treating risk as a static compliance exercise, the framework follows a lifecycle-oriented progression: first classifying potential risks, then assessing their relevance and severity in context, implementing proportionate mitigation measures, and finally ensuring ongoing oversight and accountability as systems evolve.

Throughout the module, risk management is treated as an iterative and context-dependent process, recognising that startups operate under resource constraints, rapidly changing technical architectures, and diverse deployment environments. Inclusive risk planning is integrated as a cross-cutting consideration, ensuring that risk decisions account for varied user capabilities, access conditions, and potential unintended impacts.



The AI Lifecycle

The AI model and product lifecycle can be summarised into four main stages. It should be noted that each stage comprises numerous detailed activities further specified in ISO/IEC 5338:2023[80] on AI system life cycle processes but for illustration purposes this Toolkit groups them into four high-level stages: (1) Planning and Design, (2) Data Collection and Processing, (3) Model Development and Evaluation, and (4) Model Deployment and Inference, supported by cross-cutting risk prioritisation and documentation throughout.

01

Stage 1: Planning and Design

The planning and design stage defines what the AI system is intended to do, who it is for, and what risks must be managed. Here, teams set out to clarify the problem being solved, set system boundaries, identify potential misuse, define success and safety criteria, and determine whether human oversight is required. Decisions made here may shape all downstream technical and safety outcomes.

02

Stage 2: Data Collection and Processing

This stage involves sourcing, cleaning, labeling, transforming, and storing the data used to train or configure the model. These activities may be iterative across the models lifespan since model training, evaluation and inference require data points continuously. It is at this stage that teams should assess data quality, representativeness, bias risks, privacy considerations, and security controls. Poor data practices at this stage often lead to downstream safety and fairness issues.

03

Stage 3: Model Development and Evaluation

In this stage, AI models are selected, trained, tuned, and tested. Evaluation should go beyond accuracy to include robustness, bias testing, edge-case analysis, and stress testing. At this stage teams determine whether the system meets predefined safety and performance thresholds before deployment.

04

Stage 4: Model Deployment and Inference

This stage covers integrating the model into production systems and exposing it to real users. It includes configuring guardrails, access controls, monitoring, logging, and oversight mechanisms. Real-world risks materialize here, making monitoring and incident response critical.

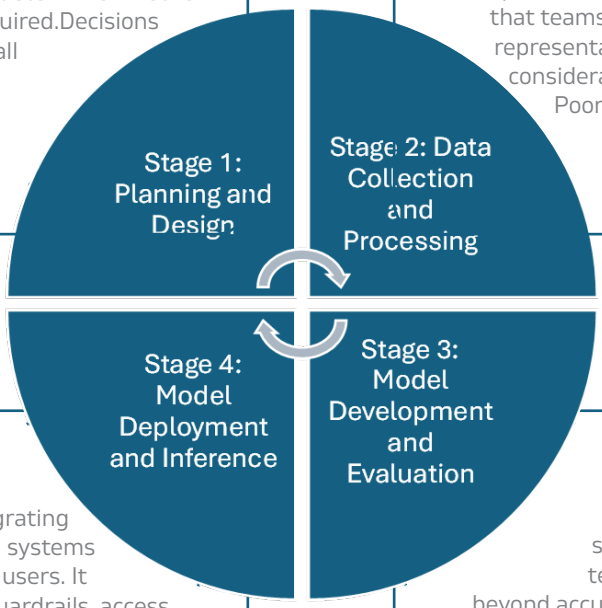
80 ISO/IEC 5338:2023, Information technology — Artificial intelligence — AI system life cycle processes https://www.iso.org/standard/81118.html accessed 10 March 2026.

Cross cutting: Risk Prioritisation & Documentation

Across all stages, risks should be identified, assessed, tracked, and updated. Teams are required to document decisions, maintain version control, monitor emerging risks, and record incidents. This ensures traceability, accountability, and continuous improvement rather than one-time compliance.

The planning and design stage defines what the AI system is intended to do, who it is for and what risks must be managed. Here, teams set out to clarify the problem being solved, set system boundaries, identify potential misuse, define success and safety criteria and determine whether human oversight is required.Decisions made here may shape all downstream technical safety outcomes.

This stage involves sourcing, cleaning, labeling, transforming and storing the data used to train or configure the model. Theses activities may be iterative across the model's lifespan since model training, evaluation and inference require data points continuously. It is at this stage that teams should assess data quality, representativeness, bias risks, privacy considerations and security controls. Poor data practices at this stage often lead to downstream saftety and fairness issues



This stage covers integrating model into production systems and exposing it to real users. It includes configuring guardrails, access controls, monitoring, logging and oversight mechanisms. Real-world risks materialize here, making monitoring and incident response critical.

This stage,AI models are selected, trained, tuned and tested. Evaluation should go beyond accuracy to include robustness, bias testing, edge-case analysis and stress testing. At this stage teams determine whether the system meets predefined safety and performance thresholds before deployment

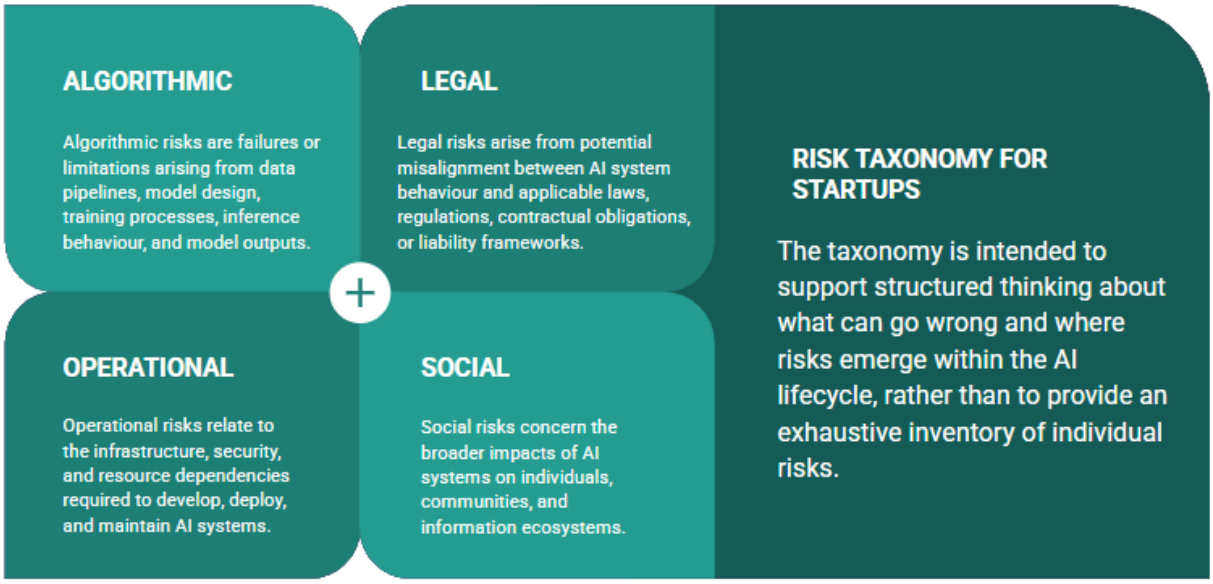
Cross Cutting Risk Prioritisation and Documentation

Across all stages, risks should be identified, assessed, tracked and updated. Teams are required to document decisions, mantain version control, monitor emerging risks and record incidents. This ensures traceability, accountability and continous improvement rather than one-time compliance

AI Risk Taxonomy for Startups

This section introduces an AI risk taxonomy that classifies common categories of risk relevant to AI systems across algorithmic, legal, operational, and social dimensions. The taxonomy is intended to support structured thinking about what can go wrong and where risks emerge within the AI lifecycle, rather than to provide an exhaustive inventory of individual risks. The risks are synthesized from established risk taxonomies and standards such as IBM Risk Atlas[81], MIT Risk Registry[82], the NIST AI Risk Management Framework (AI RMF 1.0)[83] and complementary research on privacy/data risks and human-centred harms[84,85].

The matrix below summarises common AI risks across four domains, algorithmic, legal, operational, and social mapped to key stages of the AI lifecycle: data, training, inference, and outputs. It helps startups quickly locate where risks typically originate and where they surface in user-facing behaviour. The same issue may appear across multiple domains (e.g., biased data can trigger social harms and legal exposure). Use the matrix as a navigation aid; the subsections that follow provide detailed risk statements, examples, and the tools used to assess each category.



81 F Bagehorn and others, 'AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources' (2025) arXiv preprint arXiv:2503.05780 https://arxiv.org/abs/2503.05780 accessed 10 March 2026.

82 AK Saeri and others, 'Mapping AI Risk Mitigations: Evidence Scan and Preliminary AI Risk Mitigation Taxonomy' (2025) arXiv preprint arXiv:2512.11931 https://doi.org/10.48550/arXiv.2512.11931 accessed 10 March 2026.

83 R Schwartz and others, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (NIST Special Publication 1270, National Institute of Standards and Technology 2022) https://doi.org/10.6028/NIST.SP.1270 accessed 10 March 2026.

84 G Abercrombie and others, 'A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms' (2024) arXiv preprint arXiv:2407.01294 https://doi.org/10.48550/arXiv.2407.01294 accessed 10 March 2026.

85 G Billiris, A Gill and M Bandara, 'Privacy in the Age of AI: A Taxonomy of Data Risks' (2025) arXiv preprint arXiv:2510.02357 https://doi.org/10.48550/arXiv.2510.02357 accessed 10 March 2026.

Domain	DATA	TRAINING	INFERENCE	OUTPUTS
Algorithmic	Risks start when data is biased, missing key groups, poorly labelled, or its origin and suitability are unclear, so the system learns patterns that don't reflect real users.	Risks arise if the model is trained to optimise the wrong goal, becomes unstable, amplifies bias over time, or memorises sensitive training data.	Risks occur when real-world inputs or attackers manipulate the system (including prompt attacks), or when privacy can be inferred from model behaviour even if data handling seems compliant.	Risks show up as incorrect or unsafe outputs (including hallucinations), low explainability, toxic content, or leakage of confidential or copyrighted information.
Legal	Risks arise if personal or proprietary data is collected/ used without a valid legal basis, proper consent/notice, licensing rights, or compliant cross-border transfer arrangements.	Risks arise if training uses data beyond its permitted purpose, embeds personal/copyrighted material into model weights, or lacks documentation needed to defend decisions.	Risks arise when automated decisions lack required safeguards, personal data is processed unlawfully during live use, or accountability is unclear when models call external tools/ APIs.	Risks arise if outputs are defamatory, misleading, infringe IP, disclose confidential information, or create unclear liability for downstream harms.
Operational	Risks arise from insecure storage, weak access controls, poor data versioning/lineage, or manual handling of sensitive datasets that increases exposure or loss.	Risks arise from insecure training environments, reliance on external compute/tooling without assurance, weak reproducibility, and cost or sustainability volatility from heavy experimentation.	Risks arise from insecure endpoints (abuse/DoS/ prompt injection), third-party dependency failures, lack of rollback/failover, and runaway inference workloads that strain budgets and reliability.	Risks arise when wrong outputs cascade into downstream systems, failures are hard to trace, audit evidence is missing, and reliability issues create reputational damage.
Social	Risks arise when datasets over-represent dominant groups, exclude marginalised communities, or encode historical inequities, producing systematically worse outcomes for some people.	Risks arise when models learn stereotypes, prioritise average performance over minority groups, or reinforce bias through training objectives and feedback dynamics.	Risks arise when system performance differs by user context (language, literacy, connectivity, device), feedback loops reinforce inequality, or the system nudges/manipulates users unfairly.	Risks arise as discrimination, loss of autonomy and trust, disinformation/deepfakes, psychological harm, and wider societal impacts such as polarisation or institutional distrust.

Climate and Environmental Risks

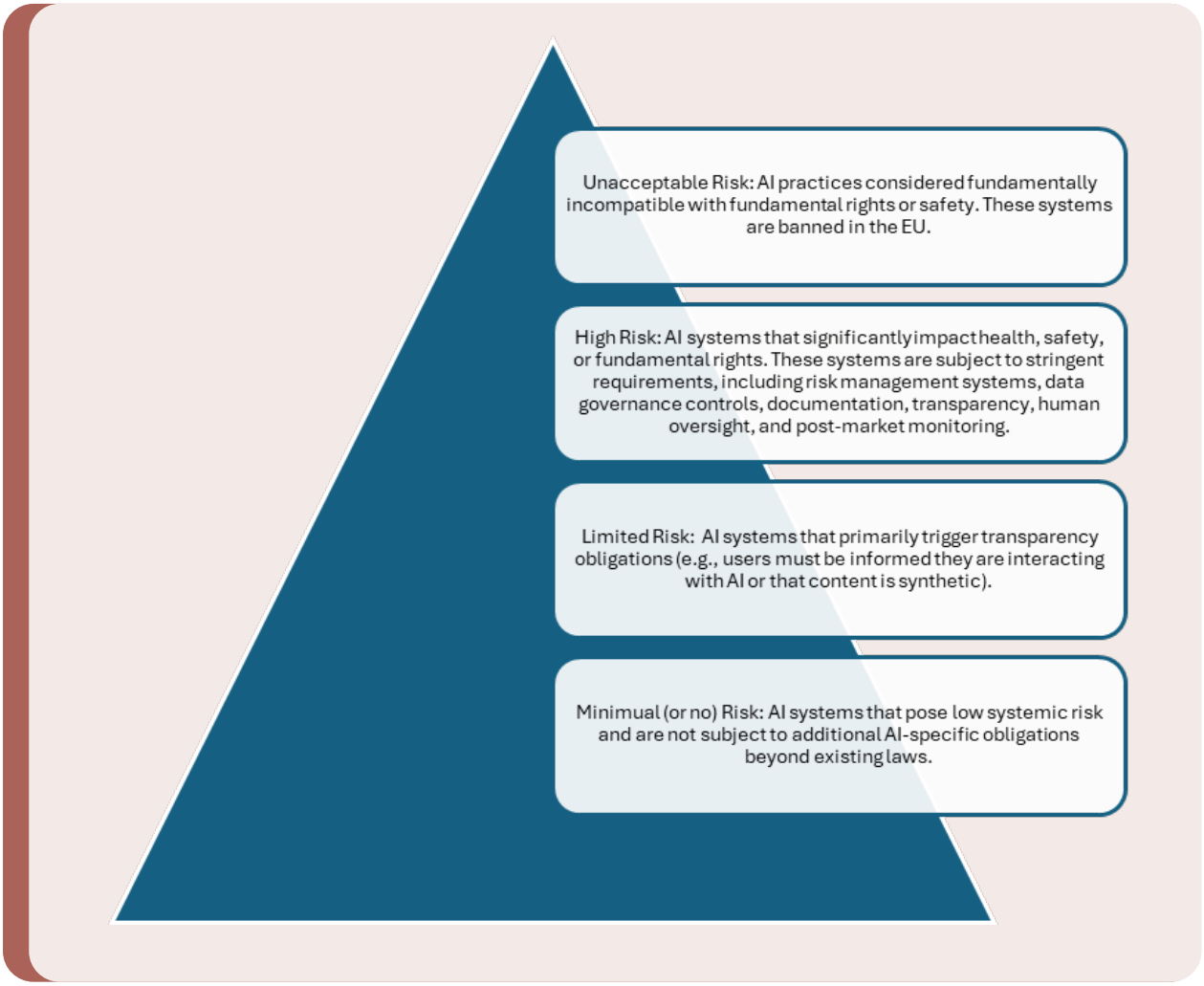
Model training and Inference also include climate and environmental risks. During model training, high energy consumption associated with training large or repeatedly fine-tuned models, leads to increased carbon emissions due to intensive use of data centre resources as well as indirect environmental impacts such as water usage for cooling and electronic waste from specialised hardware. These risks are often exacerbated by inefficient experimentation cycles and limited visibility into the environmental footprint of third-party compute providers.

During model inference, climate and environmental risks include the cumulative impact of continuous or high-volume inference workloads, especially for user-facing or agentic systems that operate at scale. Unoptimised inference pipelines, always-on services, and autonomous execution loops can result in sustained energy demand and emissions that grow over time, even when individual inference requests appear lightweight.

EU AI Act Risk Taxonomy

While our regional jurisdictions have not yet adopted a formal risk-based AI regulatory framework, startups should be aware that other jurisdictions, most notably the European Union AI Act[86], classify AI systems according to risk tiers tied to regulatory obligations.

Due to the "Brussels Effect," EU regulatory standards often influence global technology markets, procurement requirements, investor expectations, and cross-border product deployment. Startups building scalable AI systems should therefore understand how their use cases might be viewed under this framework.



86 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689 https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng accessed 10 March 2026.

I. Unacceptable Risk (Prohibited)

AI practices that are considered fundamentally incompatible with fundamental rights or safety. These systems are banned in the EU. Some examples include mass surveillance (use of biometric identification systems that categorise persons based on their biometric data to deduce sensitive identifiers such as race or sexual orientation.), manipulation of behaviour in a manner causing or likely to cause harm and social scoring.

II. High Risk

AI systems that significantly impact health, safety, or fundamental rights. These systems are subject to stringent requirements, including risk management systems, data governance controls, documentation, transparency, human oversight, and post-market monitoring. For example, AI screening tools for recruitment decisions, AI creditworthiness or loan eligibility scoring or AI for triage decisions in critical public services contexts.

III. Limited Risk

AI systems that primarily trigger transparency obligations (e.g., users must be informed they are interacting with AI or that content is synthetic). For example, Chatbots or AI interfaces that interact with people (where disclosure is needed) or synthetic or deepfake content where labelling or notice is required.

IV. Minimal Risk

AI systems that pose low systemic risk and are not subject to additional AI-specific obligations beyond existing laws. For example, spam filters, basic recommender systems for entertainment, generic productivity features (context-dependent).

Risk Assessment and Mitigation Tools and Methods

Building on the AI risk taxonomy discussed earlier in this module, this section outlines how startups can use existing tools and methods to assess AI risks in practice across the AI lifecycle. We also list various responsible AI tools in Appendix 3.

Stage	Description	Assessment and Mitigation Action
Problem Statement and Design: Is the problem we are solving appropriate for AI?	Risk assessment should begin before data collection and modelling. At the concept stage, the key objective is to confirm that the AI approach is appropriate for the problem, clarify system boundaries (advisory vs automated decisioning), and surface foreseeable harms, misuse pathways, and stakeholder impacts. This is where many "downstream" legal, social, and operational risks originate.	• Define intended purpose, users, and decision boundary for the AI system (advisory, assistive, automated). • Identify stakeholders who may be affected directly and indirectly. • Document assumptions about deployment context (geography, language, connectivity, device constraints). • Enumerate plausible misuse and failure scenarios (including malicious use). • Define go/no-go criteria for high-risk use cases based on risk assessment scoring.
Model Training & Evaluation Risks: Do training choices introduce bias, privacy, robustness, or explainability risk?	At the model training and evaluation stage, risk assessment focuses on whether the model performs reliably across relevant populations and contexts, whether it exhibits harmful disparities, whether privacy properties are acceptable (e.g., memorisation or leakage risk), and whether outputs are sufficiently explainable for the intended use.	• Define evaluation metrics beyond aggregate accuracy (e.g., subgroup performance, calibration, error concentration). • Test for unfair or disparate performance across relevant groups and operating conditions. • Assess explainability risk by confirming that model outputs can be meaningfully justified to intended users (and, where relevant, affected customers), that explanations are stable and do not expose sensitive attributes or rely on proxy reasoning. • Assess privacy-related risks relevant to the use case (e.g., memorisation, sensitive attribute inference). • Conduct interpretability and error analysis appropriate to deployment risk (especially for high-impact decisions). • Document training objectives, data versions, constraints, and known limitations for reproducibility.

Stage	Description	Assessment and Mitigation Action
Data Risks: Do our data sources introduce bias, privacy, or legal risk?	Risk assessment at the data stage focuses on identifying risks related to data quality, representativeness, consent, provenance, and reuse.	• Create a data inventory (sources, owners, sensitivity, retention, transfer, versioning, and update cadence). • Assess provenance and legal basis/ rights for collection and reuse (including third-party datasets). • Check and document representativeness, missingness, labeling quality, and known bias risks.
Deployment & Inference Risks: Will the system behave safely and securely in real-world operation?	At deployment and inference, risks often emerge that are not visible during training—particularly for systems exposed to user inputs, adversarial behaviour, complex integrations, or changing environments. Risk assessment here focuses on runtime safety, security, dependency risk, and operational resilience.	• Identify runtime threat surface (prompt attacks, injection/jailbreak patterns, extraction risks, endpoint abuse). • Review dependencies (third-party APIs, model providers, toolchains) and define failure modes/contingencies. • Establish minimum runtime safeguards (authentication, rate limits, access control, sandboxing where relevant). • Define escalation protocols for incidents and unexpected behaviour. • Confirm that the deployment context assumptions (users, languages, connectivity, device types) remain valid
Cross-cutting Risk Prioritisation & Documentation: Can we score risks consistently and produce audit-ready evidence?	Across all lifecycle stages, organizations need a practical mechanism to convert "observations" into prioritised risks, assign ownership, and preserve evidence.	• Maintain a risk register with consistent risk scoring (impact × likelihood)[87] owners, due dates, and status. • Link risk entries to evidence artifacts (data documentation, evaluation results, security reviews, impact assessments). • Define reassessment triggers (new data source, retraining, new geography/user group, major feature changes). • Establish a minimum documentation pack for external stakeholders (partners, regulators, clients).

87 R Graves, 'Qualitative Risk Assessment' (2000) 14(10) PM Network 61 https://www.pmi.org/learning/library/qualitative-risk-assessment-cheaper-faster-3188 accessed 10 March 2026..

Monitoring, and controls: Practical tools for continuous management of AI Risk

AI risks can be managed by combining preventive controls (before deployment) with detective controls (during use) and response controls (when failures happen). In practice, organizations should choose tools based on (a) system criticality (advisory vs automated decisioning), (b) data sensitivity, and (c) team capacity and budget.

A simple selection rule:

Self-hosted/open-source is often best when you need data control, lower ongoing costs, and can run infrastructure.

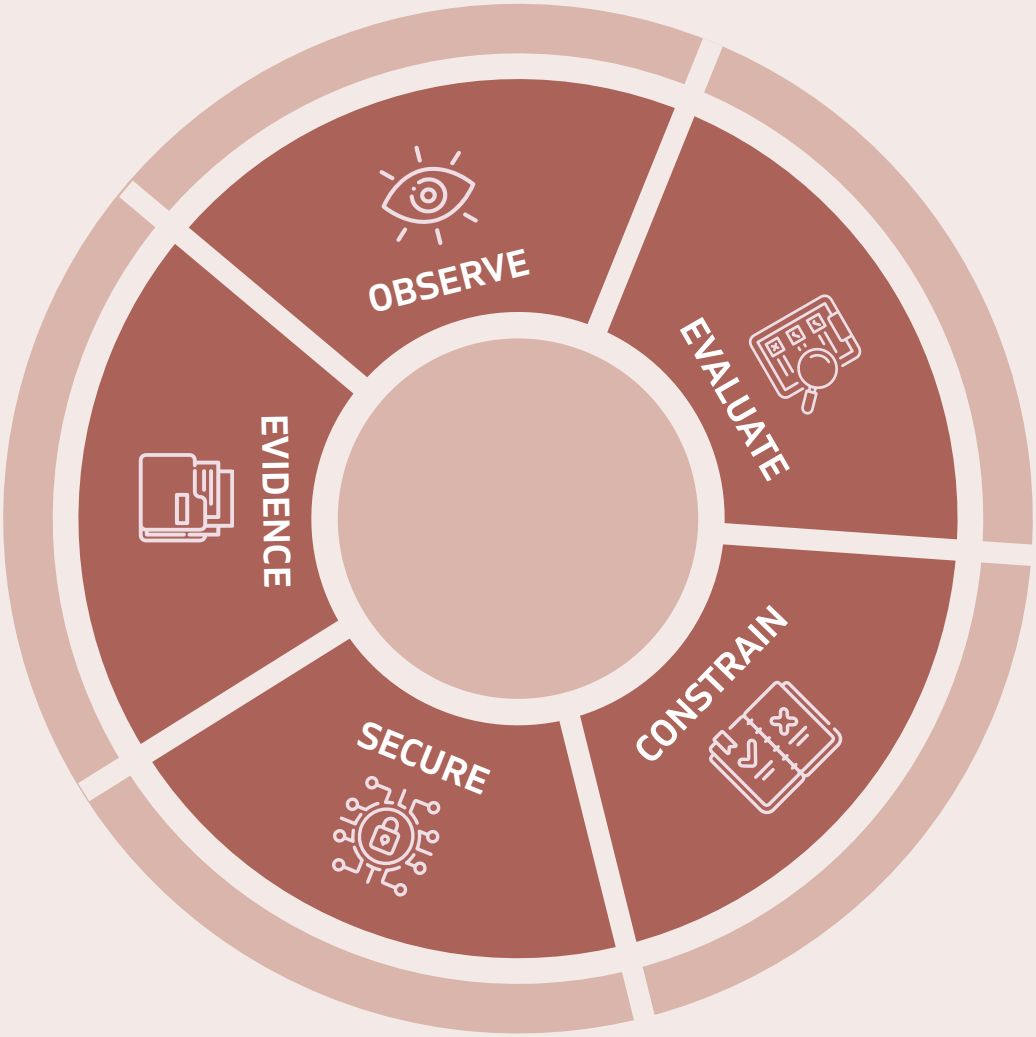
Managed services are often best when you need speed, support, and standardised workflows, especially when selling to enterprises.

Across both flavours, you should select controls proportionately based on risk assessment outcomes (impact × likelihood), the decision boundary (advisory vs automated), data sensitivity, and the deployment context. But at the bare minimum you should start with a small set of controls that create immediate safety and accountability along the following lines:

- **Traceability (being able to reconstruct what happened and why)**: You can always tell which data, model, prompts, and settings produced a decision or output so you can explain it, audit it, and fix it.
- **Release gates (a formal go/no-go check before updates go live):** you do not ship a model/prompt update without basic evaluation evidence.
- **Runtime containment (built-in limits that prevent unsafe behaviour during real use):** you constrain high-risk behaviours for user-facing AI (especially GenAI).
- **Security baseline (minimum protections and checks against common attacks):** you test for common exploit paths (especially prompt injection for LLM apps).
- **Incident readiness (preparedness to respond, recover, and learn from failures):** you can pause/rollback and run a root-cause analysis when failures occur.

To operationalize continuous management of AI risk we propose a cyclic 5-prong process consisting of the following phases: observe, evaluate, constraint, secure and evidence. Within each step we provide a set of tools that can be ingrained within AI development and deployment workflows

A note on traditional ML vs GenAI, they share the same control logic even though GenAI introduces new failure modes (e.g., prompt injection and jailbreaks). The control categories remain the same:



For GenAI, observability often emphasises tracing and prompt|output analysis, and security emphasises LLM-specific threat patterns (e.g., prompt injection).

Observe

Under observation and monitoring the aim is to detect regressions, drift, anomalies, and unexpected behaviour.

- WhyLabs AI Control Center[88] provides observability/monitoring workflows (including drift and cohort-style analysis) designed for AI systems.
- Arize Phoenix[89] is positioned as an open-source tracing and evaluation platform, widely used for LLM tracing and troubleshooting in AI applications.
- Langfuse[90] provides open-source LLM observability/tracing and can be self-hosted.
- Evidently AI[91] available as an open source self hosting tool or commercially hosted too supports evaluation, testing, and monitoring for ML and LLM systems (useful when you prefer a library-first approach and integrate into your own pipelines).

If you are primarily deploying LLM apps, prioritise tracing (Phoenix/ Langfuse). If you operate predictive ML models, prioritise drift/ performance monitoring (WhyLabs/Evidently)

Evaluate

The aim of evaluation is to prevent regressions by making evidence a prerequisite for deployment. Tools like MLflow[92] provide automation workflows for experiment tracking, model versioning/registry, and lifecycle management; it explicitly positions itself as supporting both traditional ML workflows and GenAI app/agent workflows. Startups would benefit from a single system of record (like MLflow) because it enables disciplined releases without requiring a large MLOps team.

Constrain

AI constraints are usually implemented through a combination of policy and rules enforcement, model behaviour constraints, and workflow constraints that can be built within the model stack.

At the model training layer, developers can enforce "common-sense" behaviour using shape constraints, especially where predictability and fairness are important (credit scoring, pricing, risk scores). Models such as XGBoost support monotonic constraints so specific features can be forced to only increase or only decrease predictions, helping align model behaviour with domain logic[93]. LightGBM similarly exposes a monotone_ constraints parameter, enabling monotonic relationships to be enforced

88 https://docs.whylabs.ai/docs/
89 https://phoenix.arize.com/
90 https://langfuse.com/
91 https://www.evidentlyai.com/
92 https://mlflow.org/docs/latest/
93 https://xgboost.readthedocs.io/en/latest/tutorials/monotonic.html

during training[94]. Where developers want more explicit constrained modelling patterns (monotonicity, calibration layers, and interpretable constrained structures), TensorFlow Lattice provides constrained and interpretable models and tutorials specifically demonstrating monotonicity "shape constraints" and ethically motivated constraint use[95].

At the decision layer, developers can use policy and rules engines to enforce "must or never" conditions around a model's output (for example: eligibility prerequisites, caps, mandatory manual review for certain outcomes, or jurisdiction-specific restrictions). Open Policy Agent (OPA) is an open-source policy engine that externalises decision logic ("policy-as-code") so applications can ask a policy service whether an action is allowed. In more business-rules-heavy environments, tools such as Drools can be used as a rule engine or a Business Rules Management System (BRMS) that evaluates facts against rules and decision models, enabling teams to maintain auditable decision logic outside the ML model. For lighter-weight, event-driven use cases, durable_rules provides a polyglot rules micro-framework that can coordinate and reason over event streams, useful when developers want simple "if-then" controls without a full BRMS stack.

Still within the decision layer developers can limit what automated decisions trigger in downstream systems (for example: step-up verification instead of hard rejection; two-person approval for high-impact actions; caps on automated account blocks per hour). This is where workflow engines and durable execution platforms are practical. Temporal is positioned as an open-source platform that provides "durable execution" so workflows can reliably run to completion despite failures, useful for building controlled, multi-step decision processes around ML outputs (including retries, timeouts, and long-running approvals)[96]. For organizations that want process modelling and orchestration tools, commercial tools like Camunda positions itself as a process orchestration platform, commonly used to design and automate structured workflows (e.g., approval chains, reviews, escalation paths) around automated decisions[97].

With regards to generative AI guardrails reduce the chance that unsafe, non-compliant, or malformed outputs reach users.

- **NeMo Guardrails (OSS)** is an open-source Toolkit for adding programmable guardrails to LLM conversational applications[98].
- **Guardrails (guardrails-ai, OSS)** runs input/output guards to detect and mitigate risks and supports structured output generation[99]. Alternatively, Guardrails AI (managed service) positions itself as a managed way to deploy guardrails with observability and customisation[100].

Guardrails should be applied selectively to highest-risk pathways (e.g., medical/financial guidance, eligibility decisions, sensitive data handling, tool-using agents).

94 https://lightgbm.readthedocs.io/en/latest/Parameters.html
95 https://www.tensorflow.org/lattice/overview
96 https://docs.temporal.io/ai-cookbook/human-in-the-loop-python
97 https://camunda.com/solutions/process-automation-platform/
98 https://github.com/NVIDIA-NeMo/Guardrails
99 https://github.com/guardrails-ai/guardrails
100 https://www.guardrailsai.com/

Secure

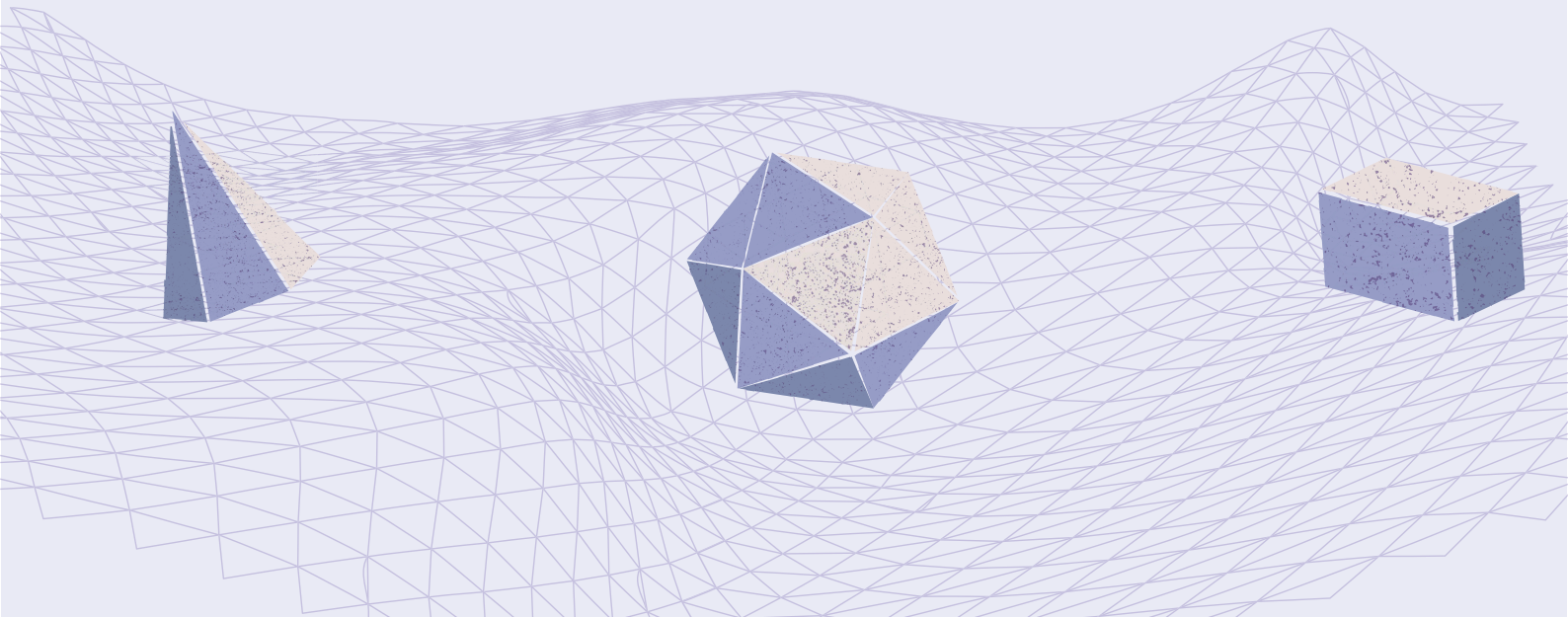
Security controls aim to identify exploitable weaknesses and reduce the threat surface (especially for user-facing GenAI applications). Tools such as OWASP Top 10 for LLM Applications provides a practical checklist of common LLM security risks such as prompt injection and insecure output handling[101]. Additionally, garak (OSS) is an LLM vulnerability scanner designed to probe ways LLMs or dialog systems fail (including jailbreak/prompt-injection style tests)[102]. You can treat OWASP LLM Top 10 as the baseline for threat modelling and use tools like garak to operationalise "pressure testing" before launch and before major releases.

Evidence

Compliance automation and evidence management reduce the overhead of audit readiness and evidence collection when customers require formal compliance. Proprietary tools like Vanta and Sprinto automate compliance and evidence management around SOC 2/ISO-related compliance automation around integrations and automated testing/evidence collection[103,104]. These tools are most justified when you sell to enterprise/regulatory customers that require a formal control environment; early-stage startups can start with a lightweight evidence pack and upgrade later. Tools like VerifyWise a commercial tool for AI governance offer starter options[105] or alternatively opt for open-source self hosting option available on Github[106].

101 https://owasp.org/www-project-top-10-for-large-language-model-applications/
102 https://github.com/NVIDIA/garak
103 https://www.vanta.com/
104 https://sprinto.com/
105 https://verifywise.ai/pricing
106 https://github.com/bluewave-labs/verifywise

04
Appendices



Appendix 1: Case Study: AI Driven Alternative Credit Scoring for MSMES in Kenya

Scenario 1 — Concept stage: "Should we really use AI here?"

A Kenyan fintech SME ("JiraniCredit") provides small working-capital loans to micro-merchants and youth-led businesses. They are considering building an AI-Driven alternative credit scoring based on mobile money and digital transaction behaviour. The approach is inspired by the market context where products like Fuliza extend short-term liquidity through M-PESA and national initiatives like the Hustler Fund seek to expand access to credit through mobile channels.

Observe	Evaluate (checklist)	Constrain	Secure	Evidence (minimum artefacts)
Currently our loan decisions are manual and inconsistent across staff. The business wants faster approvals and fewer defaults, but the customer base includes informal traders with irregular cashflows, patchy smartphone access, and mixed digital literacy.	[1.1] Intended purpose/users/boundary: Purpose is "faster lending decisions," but the boundary is unclear (recommendation vs automated approval/decline). [1.2] Stakeholders: They initially list only customers and credit officers—missing indirect stakeholders (family guarantors, customer support, collections agents, regulators, and "near-miss" applicants). [1.3] Deployment context: Assumptions are incomplete: language (English/Kiswahili), connectivity, feature phone users, and county-level differences not documented. [1.3] Failure and misuse: They have not considered gaming behaviour (SIM swap fraud, synthetic transaction patterns), exclusion risks, or debt distress. [1.4] Go/no-go criteria: No explicit thresholds for "too risky to automate," or triggers for escalation (complaints, unusual declines, disparity).	The team sets an initial boundary: assistive by default until evaluation evidence supports partial automation. They define out-of-scope uses: the system may not be used for "blacklisting" customers across products or sharing risk scores externally without governance.	They define what data is sensitive and restrict access to a small group, even before building models (minimum access principle).	• One-page Intended Use & Decision Boundary note. • Stakeholder map including indirect stakeholders. • Deployment assumptions (language/ connectivity/device constraints). • Failure modes & misuse scenarios list. • Go/no-go + escalation triggers draft.

Scenario 2 — Data stage: "Our data is convenient, but is it fair and lawful?"

After scoping the data they have and looking at state of the art alternative credit models, JiraniCredit plans to use transaction behaviour, airtime top-ups, device metadata, location patterns, and repayment history (where available). They also consider buying third-party enrichment data. Though this may introduce risks such as, indirect discrimination via proxy features (location/device), privacy and consent risk from sensitive personal/transaction data. Kenya's digital credit environment also has explicit expectations around consumer protection and data protection practices for digital lenders.[107]

107 Office of the Data Protection Commissioner, Guidance Note for Digital Credit Providers (December 2023) https://www.odpc.go.ke/wp-content/uploads/2024/02/ODPC-Guidance-Note-for-Digital-Credit-Providers.pdf accessed 10 March 2026.

Observe	Evaluate (checklist)	Constrain	Secure	Evidence (minimum artefacts)
The easiest available datasets are mobile money and app-usage logs. The team realises some customers operate mostly in cash, some share phones, and some transact through agents so the data may represent "digital visibility," not true creditworthiness.	2.1] Data inventory: No complete inventory yet (sources, owners, retention, transfer, update cadence). [2.2] Provenance and legal basis: Unclear basis for using device and location data and for any third-party data reuse. [2.3] Representativeness and bias: Under-coverage likely for rural youth, informal traders, feature phone users, and customers with irregular digital footprints. [2.4] Privacy exposure: PII is present (phone number/ID linkage) and there is a re-identification risk if datasets are combined.	Create a feature deny-list at data stage: exclude overly granular location and unnecessary device identifiers until there is a justified use case. Define a minimum "representation check" before training: ensure each priority segment has enough data to evaluate performance (even if imperfect). Implement "minimum necessary data" for the first model version.	Restrict dataset access; encrypt storage; separate identifiers from behavioural features (pseudonymise where practical). Define retention limits and secure transfer requirements for any third-party data.	• Data inventory and provenance notes. • PII and sensitivity map (what is personal, sensitive, high-risk). • Representativeness/ coverage check summary. • Feature allow/deny list with justification.

Scenario 3 — Model training: "It performs well overall, but not for everyone"

JiraniCredit now has a beta AI system that produces a risk band and a recommended loan limit for their customers. The beta version **automates decision making** for **faster approvals**.

Observe	Evaluate (checklist)	Constrain	Secure	Evidence (minimum artefacts)
Pilot results look "good on average," but customer support reports complaints from some market traders who are frequently recommended for very low limits despite steady business activity.	[3.1] Beyond accuracy: Metrics are mostly overall accuracy; no calibration or error concentration analysis. [3.2] Disparate performance: No subgroup testing across relevant segments (county/rural-urban, business type, device/connectivity). [3.3] Privacy properties: No assessment of memorisation/leakage risk (especially if training data includes sensitive attributes). [3.4] Explainability: Staff cannot consistently explain why a limit is low; "reason codes" are inconsistent. [3.5] Reproducibility: Training data versioning and objectives are not fully documented. [3.6] Environmental impact: Not material for this small model yet, but frequent retraining could become costly.	Add slice-based evaluation as a release requirement (not optional). Introduce abstention/ grey-zone logic during evaluation: borderline cases should route to manual review. Consider monotonic constraints for features where "common sense" should hold (e.g., stronger repayment history should not reduce score).	Lock down training environments; control who can retrain; keep training logs and artefacts.	• Evaluation report with subgroup slices and calibration. • Error analysis notes (where the model fails). • Explainability artefact (stable reason codes sm and limits). • Reproducibility pack (data versions, training objective, constraints, known limitations).

Scenario 4 — Deployment and inference: "Automation created harm"

JiraniCredit still deploys the beta system and automates approvals and declines to improve speed. This reduces manual workload but problems emerge.

Observe	Evaluate (checklist)	Constrain	Secure	Evidence (minimum artefacts)
Within weeks, declines increase sharply. Customer support reports two patterns: (i) inconsistent outcomes for similar applicants, and (ii) increased complaints from specific segments (e.g., micro-merchants operating through agents). Fraud attempts also increase (suspicious behaviour suggests gaming).	[4.1] Threat surface: Automation exposes new abuse pathways (API abuse, manipulation, synthetic behaviour). [4.2] Dependencies: Model relies on MNO and payment integrations; outages or changes affect inputs. [4.3] Minimum safeguards: Rate limits, access control, and rollback are weak. [4.4] Incident escalation: No clear incident triage, pause authority, or communications plan. [4.5] Deployment assumptions: Real usage differs (connectivity, language support, agent-mediated transactions). [4.6] Corrective actions: No structured tracking of nonconformities and fixes.	Immediate policy constraint: "**No auto-decline**." High-impact adverse outcomes require review. Apply caps and step-up verification instead of denial for certain risk signals. Add "grey-zone to human review" routing.	Implement authentication, rate limiting, monitoring for abuse, and minimum logging for audits. Add dependency failure contingencies (fallback behaviour when upstream data is missing).	• Incident logs. • Updated threat surface assessment. • Change log showing containment changes and rollback plan.

Scenario 5 — Governance, prioritisation, and documentation: "We set the system to assistive"

They have decided that the decision boundary of the system is to be assistive, that is, the system can auto-approve only low-risk cases within caps; it cannot auto-decline.

Observe	Evaluate (checklist)	Constrain	Secure	Evidence (minimum artefacts)
The product becomes slightly slower, but complaints reduce. The company now needs to ensure the fix is sustained and defensible to partners and regulators.	[5.1] Risk register: Not consistently maintained; no owners and due dates. [5.2] Evidence linkage: Evidence exists but is scattered; not linked to risks. [5.3] Reassessment triggers: Not formalised (new county, new dataset, model update). [5.4] External documentation pack: Not prepared (purpose, limitations, appropriate use). [5.5] Structured self-assessment: Not stored or linked (ALTAI/NIST/TrustArc). [5.6–5.7] Competence and awareness: Training records and policy awareness are informal. [5.8] Communications: No standard customer disclosure for AI-assisted decisions. [5.9] Document control: Versioning/retention unclear. [5.10] Internal audit: Not done.	Define "high-risk change" categories requiring review (new data source, new region, change in decision boundary). Make risk acceptance explicit—no "silent residual risk."	Control access to sensitive documentation and supplier model documentation; enforce versioning.	• A living risk register scored by impact × likelihood, with owners and deadlines. • Evidence index linking each checklist item to artefacts. • External documentation pack + customer disclosure. • One completed self-assessment linked to the risk register.

While inspired by Kenya's digital credit landscape (e.g., Fuliza-style overdraft services and government digital credit initiatives), this case is a composite example designed for learning and does not describe any single provider's proprietary scoring mode

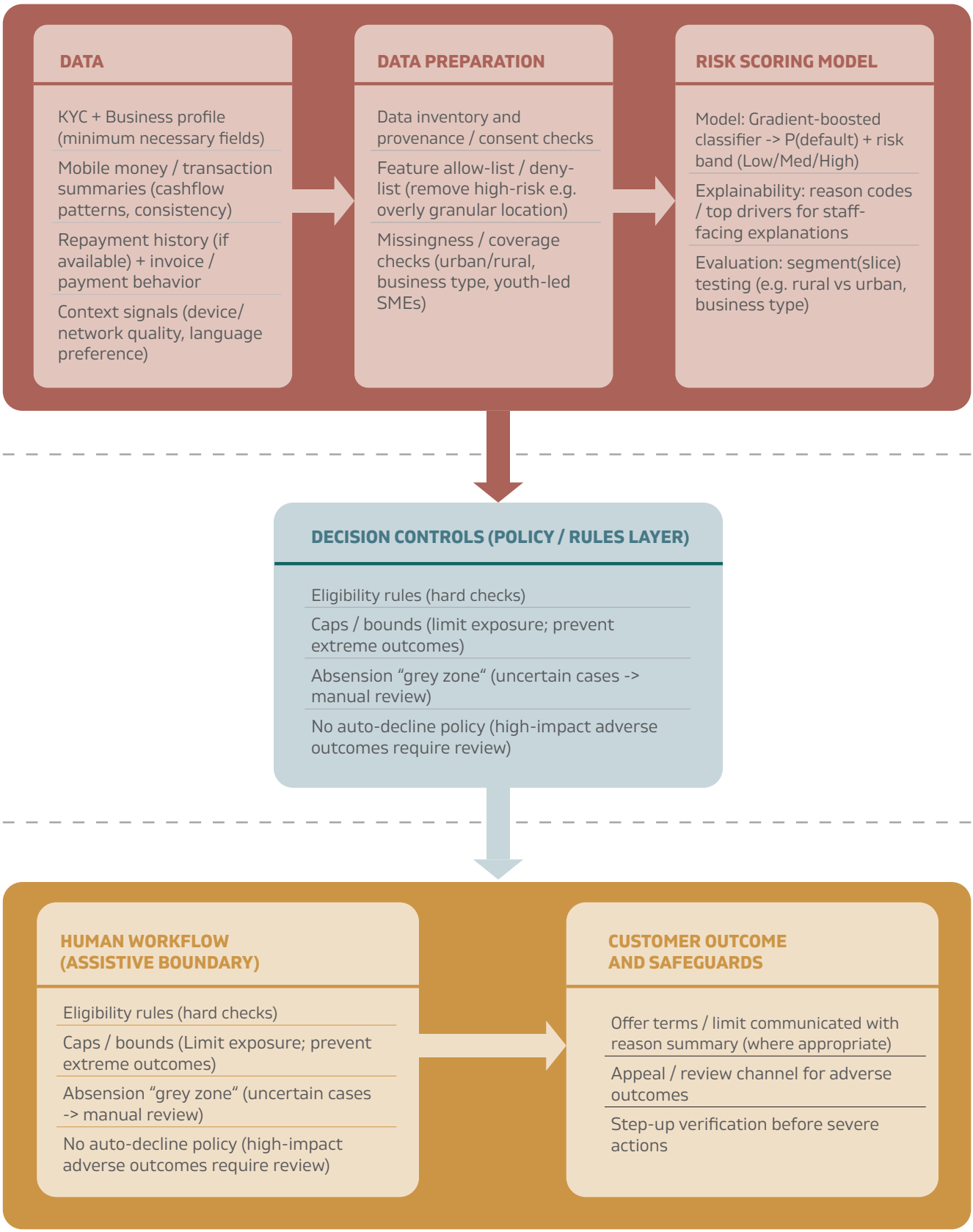
Responsible AI Success Criteria

Decisions are explainable to staff, error burden is not concentrated on specific segments (e.g., rural micro-merchants), sensitive data is minimised and protected, and customers have a review/appeal path for adverse decisions.

System Overview

Given the top risks and responsible AI success criteria, the JiraniCredit technical team developed a system that combines alternative data (KYC, transaction summaries, repayment and invoice behaviour) to generate a credit risk band and a recommended limit for MSMEs. Before modelling, data is documented and screened for provenance/consent, and features are governed through an allow-list/deny-list approach to reduce privacy exposure and proxy discrimination (e.g., avoiding overly granular location signals). A predictive machine learning model then produces a probability of default and staff-facing reason codes, and performance is checked across key segments such as rural vs urban and different business types. To prevent harm at deployment time, the model output is not applied directly. A policy/rules layer enforces eligibility checks, caps exposure, and creates an abstention "grey zone" where uncertain cases are routed to manual review. The decision boundary is intentionally assistive: low-risk cases may be auto-approved within caps, but high-impact adverse outcomes are never fully automated (no auto-decline). Finally, cross-cutting controls—observability, security safeguards, and change management with release gates and rollback—help maintain safe operation as the system scales or evolves.

JiraniCredit alternative credit scoring system with engrained RAI Principles



Legal Analysis for JiraniCredit: Compliance with Kenya's Regulatory Framework

This roadmap organizes JiraniCredit's compliance imperatives across its operational lifecycle.

Pre-Launch Phase

Compliance Step	Actionable Instructions for JiraniCredit
01 Registration with ODPC	• Register formally as a data controller and processor with the Office of the Data Protection Commissioner (ODPC). • Maintain updated registration records and renewals. • Ensure licensing prerequisites under the Digital Credit Providers Regulations are met before operations begin. • Follow ODPC Guidance Notes on mandatory registration and disclosure of processing activities.
02 Risk Assessment (DPIA)	• Conduct a Data Protection Impact Assessment (DPIA) for profiling and automated decision-making activities. • Document risks related to privacy, bias, and discrimination. • File and maintain DPIA reports for audit and regulatory review. • Align risk assessments with KEBS Code of Practice requirements for AI risk management, accountability, and documentation.
03 IP & Licensing Compliance	• Secure licenses for all training datasets and third-party components. • Ensure contracts clarify ownership of outputs and derivative works. • Maintain an IP register for audit purposes.

Launch Phase

Compliance Step	Actionable Instructions for JiraniCredit
Human Oversight in Credit Decisions	• Implement a "human-in-the-loop" review for adverse credit decisions to comply with Section 35 of the Data Protection Act. • Train staff to handle appeals and provide explanations for denials. • Document interventions to demonstrate compliance. • Ensure oversight mechanisms meet KEBS standards for fairness and explainability. • Follow ODPC Guidance Notes stressing transparency and human review in automated decisions.
Data Security Measures	• Deploy enterprise-grade security safeguards (encryption, access controls, intrusion detection). • Conduct regular penetration testing and vulnerability assessments. • Establish breach response protocols with 72-hour notification capability. • Incorporate KEBS guidance on security-by-design and AI-specific threat mitigation. • Follow ODPC Guidance Notes emphasizing secure handling of sensitive personal data.
Transparency & Customer Rights	• Provide clear, accessible privacy notices explaining data use in profiling. • Ensure consent is genuinely free and informed, avoiding complex terms. • Offer user-friendly explanations for automated decisions and maintain records of disclosures. • Follow KEBS Code of Practice requirements for transparency, documentation, and explainability. • Implement ODPC Guidance Notes stressing informed consent and clarity in customer communication.

Post-Launch Phase (Ongoing Operations)

Compliance Step	Actionable Instructions for JiraniCredit
Complaint Resolution Mechanisms 07	• Establish accessible channels for customer complaints and appeals. • Ensure timely resolution of disputes in line with Digital Credit Providers Regulations. • Document complaint handling processes and outcomes for accountability. • Align with KEBS emphasis on accountability and consumer redress. • Follow ODPC Guidance Notes requiring effective, accessible, and timely complaint resolution.
Continuous Governance & Monitoring 08	• Define governance triggers (bias discovery, new data source, regulatory change). • Conduct regular bias audits and retrain models when risks are identified. • Maintain incident/change logs linked to model versions. • Hold quarterly compliance reviews. • Ensure ongoing alignment with KEBS standards for trustworthy AI.

Why This Roadmap Matters

Legal Compliance: Meets obligations under Kenya's Data Protection Act (2019), Digital Credit Providers Regulations (2022), KEBS AI Code of Practice, and ODPC Guidance Notes.

Ethical Governance: Embeds fairness, transparency and accountability into AI-driven credit scoring.

Consumer Trust: Builds legitimacy by respecting privacy, ensuring human oversight, and providing clear redress mechanisms.

Appendix 2: Practical Application of RAI Principles

Appendix 2.1: RAI assessment checklist

1. Problem statement and design

1.1.	Have we defined the intended purpose, primary users, and decision boundary (advisory/assistive/automated)?	☐
1.2.	Have we identified all stakeholder groups who could be affected directly and indirectly (including non-users)?	☐
1.3.	Have we documented assumptions about deployment context (geography, language, connectivity, device constraints)?	☐
1.4.	Have we enumerated plausible failure modes and misuse scenarios (including malicious use)?	☐
1.5.	Do we have explicit go/no-go criteria and escalation triggers for high-risk use cases?	☐

2. Data Risks

2.1.	Do we have a data inventory (sources, owners, sensitivity, retention, transfer, versioning, update cadence)?	☐
2.2.	Have we verified provenance and legal basis (consent/rights to collect, reuse, transfer, and third-party constraints)?	☐
2.3.	Have we checked representativeness, missingness, labeling quality, and known bias risks against the intended population/context?	☐
2.4.	Have we assessed privacy exposure in data (PII presence, sensitive attributes, re-identification risk)?	☐

3. Model training and evaluation risks

3.1.	Do our evaluation metrics go beyond aggregate accuracy (subgroup performance, calibration, error concentration)?	☐
3.2.	Have we tested disparate performance across relevant groups and operating conditions (including deployment constraints)?	☐
3.3.	Have we assessed privacy properties relevant to the use case (memorisation, leakage, inference risks)?	☐
3.4.	Is the model explainable enough for the decision risk (error analysis, interpretability appropriate to impact)?	☐
3.5.	Can we reproduce results (training objectives, data versions, constraints, known limitations documented)?	☐
3.6.	Have we considered environmental impact (e.g., energy/CO_2 implications proportional to model size and retraining cadence) where material?	☐

4. Inference and deployment risks

4.1.	Have we identified the runtime threat surface (prompt injection/jailbreaks, extraction, abuse, endpoint threats)?	☐
4.2.	Have we reviewed dependencies (third-party APIs/model providers/toolchains) and defined failure modes and contingencies?	☐
4.3.	Are minimum safeguards in place (auth, rate limits, access control, sandboxing where relevant)?	☐
4.4.	Do we have escalation protocols for incidents and unexpected behaviour (triage, rollback, comms)?	☐
4.5.	Have we revalidated deployment assumptions (user environment, language, connectivity, device types) post-integration?	☐
4.6.	Do we track nonconformities/corrective actions and continual improvement?	☐

5. Cross-cutting prioritisation and documentation

5.1.	Do we maintain a risk register with consistent scoring (impact × likelihood), owners, due dates, and status?	☐
5.2.	Are risks linked to evidence artifacts (data docs, eval results, security reviews, impact assessments)?	☐
5.3.	Do we have reassessment triggers (new data source, retraining, new geography/ user group, major feature change)?	☐
5.4.	Do we maintain a minimum external documentation pack (purpose, limitations, appropriate use, known risks)?	☐
5.5.	Have we completed and stored at least one structured self-assessment/impact assessment appropriate to our risk level (e.g., ALTAI, Microsoft RAI impact assessment, NIST playbook/TrustArc template) and linked it to the risk register?	☐
5.6.	Are competence requirements defined and evidenced for people doing AI-related work (training/experience records kept)?	☐
5.7.	Are personnel aware of the AI policy, their contribution, and implications of nonconformance?	☐
5.8.	Do we have defined internal/external communication arrangements for AI risks/ impacts (what/when/with whom/how), including customer disclosures where needed?	☐
5.9.	Is documented information controlled (access, versioning, retention/disposition), including necessary external documents such as supplier model documentation?	☐
5.10.	Do we conduct internal audits at planned intervals and keep evidence of audit results	☐

Appendix 2.2: Responsible AI Compliance & Governance Toolkit

Phase 1: Foundation (Pre-Development)

Step	Objective	Guiding Audit Questions	Action Items / Evidence Required	Notes / Dependencies
1. Map Your Lawful Basis	Legally justify data processing and secure rights.	• Have you conducted a Data Protection Impact Assessment (DPIA)? • Can you account for all personal data sources? • Have you documented lawful bases (Consent, Contract, Legitimate Interest)? • If relying on Legitimate Interest, have you completed a balancing test?	• DPIA Report • Data Lineage Register • Lawful Basis Mapping Document • Legitimate Interest Assessment	File LIA with DPIA for audit trail
2. IP & Licensing Foundation	Secure ownership and usage rights.	• Do you have licenses permitting AI training/ testing? • Does your license clarify ownership of outputs? • Have you audited third-party libraries/models for compatibility?	• Data License Register • IP Ownership Agreements • Third-Party Audit Log	Ensure contracts cover derivative works

Phase 2: Design & Development

Step	Objective	Guiding Audit Questions	Action Items / Evidence Required	Notes / Dependencies
3. Define Purpose & Transparency	Define a clear purpose and communicate it.	• Is the purpose specific and meaningful (avoiding "purpose creep")? • Does your privacy notice explain AI logic, data use and evolution? • Can you provide user-friendly explanations for AI decisions?	• Draft Purpose Statement • Layered Transparency Notice • Documented Explainability Protocol	Align with consumer protection standards

Step	Objective	Guiding Audit Questions	Action Items / Evidence Required	Notes / Dependencies
4. Consent & Bias Scrutiny	Identify consent triggers and test for bias.	• Are you processing sensitive data (health, biometrics, ethnicity)? • Are automated decisions legally significant (e.g., loan denials)? • Have you tested for disparate impact across demographics? • Does your dataset represent local diversity?	• Consent Management UI • Human-in-the-loop Process • Bias Audit Report • Data Represen-tativeness Analysis	Explicit consent required for sensitive categories

Phase 3: Deployment & Operations

Step	Objective	Guiding Audit Questions	Action Items / Evidence Required	Notes / Dependencies
5. Jurisdiction & Security	Apply national laws and security by design.	• Have you identified all jurisdictions where you operate? • Have you appointed a Data Protection Officer (DPO) if required? • Have you registered with local data authorities? • Does your security framework address AI-specific risks (model inversion, adversarial attacks)?	• Jurisdiction Compliance Tracker • DPO Appointment Letter • Registration Certificates • AI Security Audit Report • PETs Feasibility Assessment	Cross-border transfers require adequacy safeguards
6. Governance & Mitigation Triggers	Establish monitoring and re-assessment triggers.	• Have you defined governance triggers (bias discovery, new data source, regulatory change)? • Do you mandate new DPIAs when material changes occur? • Do you maintain incident/change logs linked to model versions?	• - Governance Trigger Policy • Incident Response Plan • Incident & Change Log • Quarterly Review Minutes	Breach notifications must be filed within 72 hours

Appendix 3: Compiled Responsible AI Risk Assessment and Mitigation Tools along the AI Lifecycle

Phase	Tool
Problem Statement and Design: is the problem we are solving appropriate for AI?	• **Microsoft Guidelines for Human-AI Interaction** to structure early design decisions and avoid common usability and trust failures in human-AI systems. The tool consists of 18 evidence-based best practices for designing human-centered AI systems[108,109]. Organizations can utilize the HAX Workbook to prioritize which of the 18 guidelines are most critical for their specific use case, helping to allocate limited resources to the highest-impact risk areas[110]. • Google **People + AI Research (PAIR) Guidebook** to guide product scoping, user research, and human-centered design decisions that reduce risk before technical build-out[111]. (More tools for human-centered design and usability tests[112]) • When specifically building tools for automation, understanding how these tools affect the user workflows is important for successful human-computer interaction. The **Human-Machine Teaming Systems Engineering Guide** provides a guide to help systems developers design for human-computer collaboration looking at tailoring user requirements for autonomy and automation[113]. Similarly, the **Human-Centered Artificial Intelligence (HCAI)** framework clarifies how to design for high levels of human control and high levels of computer automation so as to increase human performance and understand the situations in which full human control or full computer control are necessary, and avoid the dangers of excessive human control or excessive computer control[114]

108 https://www.microsoft.com/en-us/haxToolkit/library/

109 S Amershi and others, 'Guidelines for Human-AI Interaction' in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (ACM 2019) 1 https://doi.org/10.1145/3290605.3300233 accessed 10 March 2026..

110 https://www.microsoft.com/en-us/haxtoolkit/workbook/

111 https://pair.withgoogle.com/guidebook/

112 https://digital.gov/topics/usability/

113 P McDermott and others, Human-Machine Teaming Systems Engineering Guide (MITRE Product MP180941, MITRE Corporation 2018) https://www.mitre.org/sites/default/files/2021-11/prs-17-4208-human-machine-teaming-systems-engineering-guide.pdf accessed 10 March 2026.

114 B Shneiderman, 'Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy' (2020) 36(6) International Journal of Human-Computer Interaction 495 https://doi.org/10.1080/10447318.2020.1741118 accessed 10 March 2026.

Phase	Tool
	• **AI Risk Assessment tools** to capture system purpose, stakeholders, risk exposure, and early risk statements in a structured, repeatable format (useful for organizations that need documentation discipline early). There are several risk assessment tools such as the **NIST AI Risk Management Framework**, a voluntary, foundational guide for managing risks in AI systems, the playbook is available as spreadsheet and online tool that requires answers along 4 domains Govern, Map, Measure and Manage[115]. **TrustArc AI Risk Assessment** is another template to systematically identify, evaluate, and mitigate potential risks associated with the development, acquisition or procurement, deployment, and operation of artificial intelligence systems based on NIST AI RMF and EU AI Act[116]. This template serves as a guide for organizations to proactively address ethical, safety, security, and societal concerns related to AI technologies. This template is a form of self-assessment intended to provide an initial approach to the evaluation of a trustworthy AI. The **AI Incidents Database** can help anticipate which harms can be realised in the real world by the development of AI for risk assessments[117], and the **cognitive bias inventories** and **common biases inventories** list biases from the social world that can be imported into AI models[118,119].
Data Risks: Do our data sources introduce bias, privacy, or legal risk?	• To create a data inventory, use a dataset documentation tool such as **Datasheets for Datasets[120]**, or use case specific variations such as **Datasheets for ML sensors[121]** or **Datasheets for Open Datasheets[122]** to move beyond simple spreadsheets and create a professional, audit-ready data inventory. These documents act as a "birth certificate" for your data, detailing its origins, composition, processing, limitations, and intended use to mitigate AI risk. These can be managed using low-cost tools such as spreadsheets based on template questions[123] or use specialized governance software platforms such as Hugging Face[124].

115 https://airc.nist.gov/airmf-resources/playbook/

116 https://trustarc.com/resource/ai-risk-assessment/

117 https://incidentdatabase.ai/

118 https://www.visualcapitalist.com/50-cognitive-biases-in-the-modern-world/

119 N Swinger and others, 'What Are the Biases in My Word Embedding?' in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (ACM 2019) 305 https://doi.org/10.1145/3306618.3314270 accessed 10 March 2026.

120 T Gebru and others, 'Datasheets for Datasets' (2021) 64(12) Communications of the ACM 86 https://doi.org/10.1145/3458723 accessed 10 March 2026.

121 Stewart, M., Warden, P., Omri, Y., Prakash, S., Santos, J., Hymel, S., ... & Janapa Reddi, V. (2023). Datasheets for Machine Learning Sensors. arXiv preprint arXiv:2306.08848.

122 AC Roman and others, 'Open Datasheets: Machine-Readable Documentation for Open Datasets and Responsible AI Assessments' (2023) arXiv preprint arXiv:2312.06153 https://doi.org/10.48550/arXiv.2312.06153 accessed 10 March 2026.

123 https://www.microsoft.com/en-us/research/wp-content/uploads/2022/07/aether-datadoc-082522.pdf

124 https://huggingface.co/docs/datasets/dataset_card

Phase	Tool
	• The **AIDRIN – AI Data Readiness Inspector** is a multipurpose python based framework that evaluates the readiness of datasets for AI and machine learning workflows across 6 categories: data quality, understandability and usability, structure and organization, data governance, impact of data on AI and fair and unbiased data[125]. This framework is particularly valuable to embedded in AI workflows before model training begins to understand the vulnerabilities in the dataset and whenever datasets are updated or repurposed[126]. • **Deon Data Ethics Checklist** as a command line rapid "stop/go" ethics gate for data sourcing and processing decisions, particularly useful when teams are moving quickly or relying on third-party datasets[127]. • **IBM Data privacy Toolkit**, a Java/Scala Library but can also be accessed via a REST APIs[128] and Microsoft Presidio SDK[129] are Toolkits for identifying the type of data to be analysed, provides masking and anonymisation for protecting sensitive data and a privacy risk identification component that identifies risks when they appear past a specified threshold. However, these automated detection mechanisms may not work for low resource languages when detecting and protecting PII data like low resource language names. However, there are tools using LLMs to help in the data collection and annotation for tasks like NER and text classification such as Prodigy[130]. • Other technical tools include to support decision making include **Know Your Data** (data exploration and documentation approach) to support systematic understanding of dataset composition and potential blind spots before training[131].

125 K Hiniduma, JL Bez, R Madduri and S Byna, 'AIDRIN: A Comprehensive Toolset for Automating Data Preparation for AI' (SC25 Research & ACM SRC Posters, 2025) https://sc25.conference-program.com/presentation/?id=int_post104&sess=sess539 accessed 10 March 2026

126 https://test.pypi.org/project/aidrin/

127 https://deon.drivendata.org/

128 https://github.com/IBM/data-privacy-toolkit

129 https://microsoft.github.io/presidio/

130 https://prodi.gy/docs/large-language-models

131 https://knowyourdata.withgoogle.com/

Phase	Tool
Model Training & Evaluation Risks: Do training choices introduce bias, privacy, robustness, or explainability risk?	• ***Fairness Indicators*** to measure and monitor model performance across different slices of data and user groups. ¤ **IBM AI Fairness 360 (AIF360) Toolkit** hosted by the Linux foundation under the Trusted-AI Toolkit[132] provides a comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models[133]. ¤ **Aequitas: Bias Auditing & "Correction" Toolkit** provides a framework for auditing predictors of and experimenting with "correcting biased models" using Fair ML methods in binary classification settings[134]. ¤ **What-If Tool** for interactive error exploration, counterfactual analysis, and scenario testing to understand model behaviour beyond summary metrics[135]. ¤ **Diverse Counterfactual Explanations (DiCE)** for ML provides a framework for implementing counterfactual explanations for binary and multiclass classifiers[136]. ¤ **TensforFLow fairness indicators** library that enables computation of commonly-identified fairness metrics for binary and multiclass classifiers[137]
	• ***Model Explainability*** tools to support interpretability workflows during training and evaluation, enabling teams to inspect model explanations, errors, and fairness-related behaviour. ¤ **Learning Interpretability Tool (LIT)** under the Google PAIR-code Toolkit provides a framework to visually interrogate model outputs that is identity which features impact model performance the most and perform what-if analysis by analyzing performance against changes in sensitive attributes. The framework supports text, image, and tabular data and can be run as a standalone server, or inside of notebook environments[138].

132 https://github.com/Trusted-AI
133 https://github.com/Trusted-AI/AIF360
134 https://github.com/dssg/aequitas
135 https://pair-code.github.io/what-if-tool/
136 https://github.com/interpretml/DiCE
137 https://research.google/blog/fairness-indicators-scalable-infrastructure-for-fair-ml-systems/
138 https://github.com/PAIR-code/lit

Phase	Tool
	¤ **AI Explainability 360 (AIX360)** is a framework also under the Trusted-AI Toolkit that provides algorithms that cover different dimensions of explanations along with proxy explainability techniques such as **LIME (Local Interpretable Model-agnostic Explanations)**[139] and SHAP **(SHapley Additive exPlanations)**[140] for tabular, text, images, and time series data[141]. ¤ Uncertainty quantification such as **IBM Uncertainty Quantification 360**[142] and **EQUINE**[143] to quantify trust of labels in supervised machine learning problems. ¤ **Explainer dashboard** package provide a codebase for deploying a dashboard web app the explains the working of a machine learning model to communicate model explainability[144].
	• ***Model robustness*** testing to determine how well models hold up to vulnerabilities and adversarial threats such as evasion, poisoning. The list of failure models in machine learning can be found in this article[145]. ¤ **AdversarialGLUE** is designed to evaluate the robustness of natural language understanding models against adversarial examples[146]. ¤ **IBM Adversarial Robustness Toolbox (ART)** by the Linux Foundation AI and Data Foundation provides frameworks for evaluating Model robustness to threats from Evasion, Poisoning, Extraction, and Inference and supports all popular machine learning frameworks[147]. Other frameworks have been built on top of this framework such as **Armory Testbed** that provides a python framework for running scalable evaluations of adversarial and defences techniques that include orchestration pipelines and result presentations[148].

139 https://github.com/marcotcr/lime
140 https://shap.readthedocs.io/en/latest/index.html
141 https://github.com/Trusted-AI/AIX360
142 https://github.com/IBM/UQ360
143 https://github.com/mit-ll-responsible-ai/equine
144 https://github.com/oegedijk/explainerdashboard
145 https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning
146 https://adversarialglue.github.io/
147 https://github.com/Trusted-AI/adversarial-robustness-toolbox?tab=readme-ov-file
148 https://github.com/twosixlabs/armory-library

Phase	Tool
	¤ **Qdata Research Trustworthy AI Toolkit** contains a suite of techniques to automatically evaluate and improve the adversarial robustness of deep NLP frameworks such as frameworks or testing, detecting and evading adversarial attacks for deep learning classifiers[149]. ¤ For Generative AI there is **BenchLLM framework** for running and evaluating LLM models within the development and deployment pipelines[150]. For agentic LLM workflows there is **AgentBench** that assesses reasoning and decision-making skills though this is an area of ongoing research[151,152]. Other GEN AI evaluation and benchmarking frameworks include Project Moonshot PromptBench ¤ **CheckList** provides code for behavioral testing of NLP models and includes frameworks for generating data, robustness testing with data perturbations, minimum functionality testing, invariance testing, and directed expectation testing[153]. ¤ **DecodingTrust** aims at providing a thorough assessment of trustworthiness in GPT models along the lines of toxicity, stereotypes and bias, adversarial robustness, privacy, machine ethics and fairness[154]. ¤ **Guardrails AI** provides a library of orchestrated, rigorously tested guardrails. For evaluating unsafe and unethical LLM risks[155]. ¤ **HotpotQA** QA dataset featuring natural multi-hop questions provide example QA datasets for evaluating LLMs[156]. ¤ **NeMo-Aligner** is a Toolkit for end-to-end LLM model alignment for safety and harm reduction[157].

149 https://qdata.github.io/qdata-page/categories/AItrust/

150 https://benchllm.com/

151 https://github.com/THUDM/AgentBench

152 X Liu and others, 'AgentBench: Evaluating LLMs as Agents' (2023) arXiv preprint arXiv:2308.03688 https://doi.org/10.48550/arXiv.2308.03688 accessed 10 March 2026

153 https://github.com/marcotcr/checklist

154 https://decodingtrust.github.io/

155 https://www.guardrailsai.com/

156 https://hotpotqa.github.io/

157 https://github.com/NVIDIA/NeMo-Aligner?tab=readme-ov-file

Phase	Tool
	• ***Model Remediation*** to diagnose and reduce observed harms (e.g., subgroup disparity) during model development—useful because it links assessment outcomes to actionable interventions. ¤ **Tensorflow Model Remediation** is a python library that contains techniques for model remediation in particular when trying to achieve equality of opportunity (model performs well for all values of a sensitive attribute) or counterfactual fairness (model performance does not change between counterfactual pairs)[158].
	• **Google Assess Privacy Risk** tool to identify and reason about privacy vulnerabilities associated with trained models (particularly relevant where personal or sensitive data is involved)[159]. • Environmental Impact assessment tools measures and reports on model performance with regards to pollution. **Machine Learning's CO2 Impact** assess and report environmental impact assessment tools can be a starting point but is not actively maintained[160]. • Available Toolkits can be bundled within evaluation workflows such as **Microsoft Responsible AI Toolbox** and **Infosys Responsible AI Toolkit** as integrated assessment suites that assess models across interpretability, fairness, error analysis, and responsible AI checks into more operational tooling. • Documentation frameworks such as model cards are important frameworks for understanding, sharing, and improving machine learning models. This can be a single tool that is accessible to diverse stakeholders impacted by the machine learning model to understand and question the underlying assumptions. **Hugging face** provide a **model card** creating tool and guidebook to help developers document their machine learning models[161].

158 https://github.com/tensorflow/model-remediation
159 https://www.tensorflow.org/responsible_ai/privacy/guide
160 https://github.com/mlco2/impact
161 https://huggingface.co/blog/model-cards

Phase	Tool
Inference & Deployment Risks: Will the system behave safely and securely in real-world operation?	• Risk assessment frameworks such as **Microsoft AI Risk Assessment (security-oriented)** to structure AI security review across the lifecycle, including deployment controls and monitoring considerations[162]. **Google Secure AI Framework (SAIF) Risk Assessment** tool provides a checklist to assess AI risks for stronger security practices[163]. • **Alibi Detect** is a python library focused on outlier, adversarial and drift detection to cover both online and offline detectors for tabular data, text, images and time series[164]. • **Garak** is a framework whose focus is on risks that are inherent in and unique to LLM deployment, such as prompt injection, jailbreaks, guardrail bypass and text replay[165]. • **Arize Phoenix** is Open Source framework for LLM tracing and evaluation in real time during development and production [166]. • **BenchLLM framework** can be used in production as well to monitor models performance and detect regressions in production. • **Deepchecks** is a proprietary LLM evaluation platform that automates standardized performance metrics, credible auto-scoring, and streamlined version comparisons[167]. • **Deepval** is an open-source LLM evaluation framework, for evaluating and testing large-language model systems based on metrics such as G-Eval, task completion, answer relevancy, hallucination, etc., which uses LLM-as-a-judge and other NLP models[168]. The framework is also available on a paid platform Confident AI to run and visualise results[169]. • **LLM Guard** by Protect AI is a suite of tools to protect LLM applications to help the detection of harmful language, prevention of data leakage, and resistance against prompt injection attacks[170]. Similarly **NeMo Guardrails** is another open source Toolkit to program guardrails to LLM-based conversational applications[171].

162 https://raw.githubusercontent.com/Azure/AI-Security-Risk-Assessment/main/AI_Risk_Assessment_v4.1.4.pdf
163 https://blog.google/innovation-and-ai/technology/safety-security/google-ai-saif-risk-assessment/
164 https://github.com/SeldonIO/alibi-detect
165 https://docs.garak.ai/garak/overview/our-features
166 https://phoenix.arize.com/
167 https://www.deepchecks.com/llm-evaluation/framework/
168 https://github.com/confident-ai/deepeval
169 https://www.confident-ai.com/?utm_source=Github
170 https://github.com/protectai/llm-guard
171 https://github.com/NVIDIA-NeMo/Guardrails

Phase	Tool
	• **MLFlow** is an open source platform for tracking, managing, and deploying models in the machine learning lifecycle, supporting multiple ML and LLM libraries[172]. MLFlow offers automated RAG evaluation functions as part of its larger automation framework. It is similar to other tools in that it will automatically generate questions from context to supplement human-written tools, and does automated RAG response and response scoring via LLM (can be two different LLMs). MLFlow also makes it easy to put together and gather data from multiple pre-planned evaluation runs. • For GenAI-specific development, the **Responsible Generative AI Toolkit** can be used to assess safety boundaries and common failure patterns (e.g., unsafe content, ungrounded responses) during iterative prompt/model refinement[173].
Cross-cutting Risk Prioritisation & Documentation: Can we score risks consistently and produce audit-ready evidence?	• Risk scoring questionnaires to standardise prioritisation across teams and cycles. Organizations can apply the Frameworks such as the RAI Toolkit Assessment[174] or NIST Framework to workshop and prioritize data privacy and ensure AI systems are fair, secure, and legally compliant. • Model Cards for Model Reporting to communicate model purpose, performance, limitations, ethical considerations, and appropriate use in a consistent, reusable format. • SME AI governance checklists to ensure the organizational foundations exist to act on assessment findings (inventory, ownership, escalation, review cadence). • **IEEE CertifAIEd Professional Certification** (where relevant) as an external professionalisation signal and capability-building route for teams formalising Responsible AI practice[175].

172 https://mlflow.org/

173 https://ai.google.dev/responsible

174 https://rai.acqbot.com/

175 https://standards.ieee.org/products-programs/icap/ieee-certifaied/professional-certification/

Appendix 4: Training Delivery Framework

Appendix 4.1: Facilitator Pack Module 1&2

Module 1: Foundational Principles of Responsible AI

Learning Objectives:

- Understand core RAI principles and their strategic value in African markets
- Distinguish between different forms of fairness and discrimination in AI systems
- Apply human-centered design approaches that respect cultural context and Ubuntu values
- Implement transparency and explainability mechanisms for user trust
- Design contestability processes enabling users to challenge AI decisions
- Apply privacy by design principles throughout the AI lifecycle
- Establish clear accountability structures for AI governance

Topics:

- 1.1. Fairness
- 1.2. Human-Centered Values
- 1.3. Transparency and Explainability
- 1.4. Contestability
- 1.5. Data Privacy and Security
- 1.6. Accountability

Module 2: Regulatory Compliance and Governance

Learning Objectives:

- Overview of applicable data protection requirements across African jurisdictions
- Establish lawful bases for processing personal data in AI systems
- Assess and mitigate IP risks including TDM and traditional knowledge
- Implement bias detection and auditing processes for algorithmic fairness
- Deploy security measures against AI-specific threats
- Develop governance frameworks for ongoing compliance

Topics:

- 2.1. Data Protection
- 2.2. Intellectual Property
- 2.3. Algorithmic Fairness
- 2.4. Data Security

Course Introduction

Facilitator introduction: Introduce yourself and your role. Clarify that the course facilitates discussion and builds on participants' experience.

Participant introduction: Each participant shares their name, role, startup focus and one ethical challenge they face or anticipate.

Context: RAI is a strategic enabler and competitive advantage in Africa, not just compliance. It protects brand reputation, builds trust, and enables responsible scaling.

Agenda: Overview of Module 1 principles and Module 2 regulatory areas.

Supporting Materials

Handouts

Module 1: Principles Definition Cards and Case Study Exercise.

Module 2: Data Protection Concepts and IP Protection Overview using the Case study as a guide.

Methodological Support

Case study discussion guidesand reflection prompts.

Further Reading

African data protection and IP laws, global AI ethics frameworks and standards, practical implementation guides and toolkits, bias and fairness resources, African digital rights and governance materials.

Appendix 4.2: Facilitator Pack Module 1&3

Enable startup teams to understand technical fairness; what it is, how it fails, how it is measured, and why metrics can conflict and to apply a practical AI risk assessment and mitigation workflow across the AI lifecycle.

The pack includes three scenarios that the facilitator can select from and two exercises corresponding to **Module 1: Technical Fairness in AI Fundamentals** and **Module 3: Technical AI Risk Assessment and Mitigation Framework**

Learning outcomes

- Participants should be able to:
- Distinguish fairness in data vs model vs deployment context and why each matters.
- Use the toolkit to identify risks across algorithmic, legal, operational and social domains using an illustrative scenario.

Scenario Handout Option 1:
Fintech Credit: “Equal approvals, unequal harm”

Use case: Alternative credit scoring for MSMEs (Kenya)

Primary fairness tension: Demographic parity vs harm via loan terms (pricing and affordability)

Dominant risk domains: Data, Social, Legal and Algorithmic

Scenario:

JiraniCredit launches an AI-driven loan decision system for micro-merchants and youth-led businesses. After a pilot, leadership mandates that the model must achieve equal approval rates across two regions (Region A and Region B) to demonstrate inclusion to partners.

The data science team adjusts thresholds until approval rates match. However, to control portfolio risk, they introduce “risk-based pricing”: customers from Region B receive higher interest rates and shorter repayment windows, because historical defaults were higher in Region B.

Within weeks, customer support records a spike in complaints from Region B: borrowers say they were “approved but set up to fail.” Community organizsations report debt distress and “repeat borrowing” among some groups. Internally, the model dashboard shows the inclusion KPI is being met.

What you know:

- Region B has higher rates of informal, seasonal income and more agent-mediated mobile money usage.
- The model uses transaction behaviour, repayment history, and some device/location proxies.
- Appeals and human review exist, but only after a borrower defaults or complains.

Scenario Handout Option 2:
Agri GenAI Advisory: "Works in testing, fails in the field"

Use case: GenAI assistant for crop pest and disease advice via WhatsApp and SMS

Primary fairness tension: Equal quality overall vs unequal usability and access in terms of language, literacy and connectivity.

Dominant risk domains: Social, Operational and Legal - if it drifts into regulated advice.

Scenario

A startup deploys a GenAI chatbot that gives pest diagnosis and treatment advice. In internal testing, the model scores high on "accuracy" using curated English prompts and clear photos.

After launch, field partners report unequal outcomes:

- Farmers with smartphones and stable connectivity get detailed step-by-step guidance.
- Farmers using feature phones, intermittent connectivity, and low literacy receive fragmented responses or instructions they can't follow.
- Kiswahili responses are shorter and sometimes omit safety warnings found in the English version.
- Some farmers follow advice incorrectly and report crop losses.
- The team argues "the model is accurate," but partner organizsations claim the tool systematically disadvantages certain users and could widen inequality.

What you know:

- The product supports English and Kiswahili, but most test data was in English.
- Many users send partial messages ("brown spots on leaves") without photos.
- The bot occasionally suggests pesticides without emphasizing protective measures.

Scenario Handout Option 3:
HR Screening:"Fair overall, unequal errors"

Use case: CV screening and candidate ranking for entry-level roles

Primary fairness tension: Overall accuracy vs equal opportunity (false negatives)

Dominant risk domains: Data, Algorithmic, Legal and Social

Scenario:

A startup sells an AI screening tool to SMEs to shortlist candidates. The model is trained on past hiring outcomes, that is who got interviewed and hired. In validation, the model shows strong overall accuracy.

After deployment, a partner company audits outcomes and finds:

- The model rejects a significantly higher proportion of qualified candidates from the following groups:
  - Women who are returning to work after a career break.
  - Candidates from certain colleges.
  - Candidates with non-linear career paths.
- When recruiters manually review rejected applications, they discover many "good fits" are being filtered out.
- The startup's product team is under pressure to ship quickly and argues that "the model is objective" and "improves efficiency." A client hints they may terminate due to discrimination risk.

What you know:

- Training labels reflect historical recruiter decisions (which may encode bias).
- The model uses features like employment gaps, institution, and prior job titles.
- There is no formal appeal or "human review for adverse decisions" process.

Module 1: Technical Fairness in AI Fundamentals

Objective

- Identify the fairness failure in the scenario provided
- Show why one fairness metric can look good while harm persists
- Decide what to test

Instructions

- Facilitator goes through Module 1: Technical Fairness in AI Fundamentals to discuss the following topics:
- Technical fairness terms
- Outcome parity in terms of selection and approval rate
- Error parity in terms of false negatives/false positives by group

Group Discussion Guide

- Select one scenario and use it to answer the following questions:
- What is the fairness failure here (e.g. parity, error rates, calibration, terms, accessibility)?
- Which fairness metric might look "good" while the system is still harmful?
- What is the likely root cause (data, labeling, objective function, threshold, interface or deployment constraints)?
- What would you measure to confirm the harm (2–3 concrete checks)?
- What decision boundary change could reduce harm (assistive vs automated, no auto-decline, caps, manual review, disclosures)?

Personal reflections and sharing (optional 10–15 minutes)

- Invite volunteers to share personal or professional anecdotes:
- What technical fairness aspects have they faced as a company? What technical fairness aspects possess a high risk?
- How did it have an impact on you or the organization?

Module 3: Technical AI Risk Assessment and Mitigation Framework

Objective:

- To understand the data, algorithms, legal and social risks that may arise from AI implementation
- To appreciate the risks assessment and mitigation options across the AI lifecycle

Instructions:

- The facilitator goes through Module 3: Technical Risk Assessment and Mitigation Framework. A worked example is provided in the case study for the facilitator to use during the session.
- After the session the facilitator can select the worked scenario or any of the other two scenarios for the participants to work in groups discussing the 10 high-leverage checklist items below against the scenario selected.
- Participants are required to fill in the "Your notes" column. Keep answers short and practical. Then at a plenary present and discuss their answers.

Walkthrough Table (Participants fill "Your notes")

#	Checklist item	Questions to guide your discussion	Your notes (write short bullet answers)
1	1.1 Intended purpose, users, decision boundary	What is the system for? Who uses it? Is it advisory/assistive/ automated? What must it not do? Where is human oversight required?	
2	1.3 Deployment assumptions (context)	What assumptions are we making about geography, language, connectivity, device type, digital literacy, and income patterns? Which assumptions are "risky if wrong"?	
3	1.3 Failure modes & misuse scenarios	List the top 3 ways the system could fail or be misused. Include at least one malicious scenario. Who is harmed?	

#	Checklist item	Questions to guide your discussion	Your notes (write short bullet answers)
4	1.4 Go/no-go criteria & escalation triggers	What conditions would stop launch or trigger rollback? What signals trigger escalation to human review (or a pause)? Who decides?	
5	2.1 Data inventory (minimum)	What data sources will we use? Who owns them? Which are sensitive? How often do they update?	
6	2.3 Representativeness / missingness / label risk	Who is under-represented or poorly measured? What missingness patterns matter? Are labels biased (reflecting historic decisions/terms)?	
7	3.1 Beyond accuracy metrics	What must we measure beyond accuracy before shipping? Pick 3 (e.g., subgroup performance, calibration, error concentration, affordability stress).	
8	3.2 Disparate performance tests (slices)	Which 2–4 slices are non-negotiable to test? Include at least one "operating condition" slice (e.g., connectivity/device).	
9	4.1 Runtime threat surface	When exposed to real users/ APIs, what are the top 2 threats? How would we detect them? What is the first-line mitigation?	
10	4.4 Incident escalation (triage, rollback, comms)	If harm happens, who can pause the system? What is the triage flow? What do we communicate, and to whom? What evidence must we capture?	

Bibliography

Books

Catherine Regis, Jean-Louis Denis, Maria Luciana Axente & Atsuo Kishimoto , Human-Centred AI :A Multidisciplinary Perspective for Policy-Makers, Auditors and Users (2024 1st edition Chapman and Hall/CRC)

Damian Okaibedi Eke, Kutoma Wakunuma & Simisola Akintoye, Responsible AI in Africa: Challenges and Opportunities (2023 Palgrave Macmillan)

Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ. Burr, C., Aitken, M., Katell, M., Fischer, C., Wong, J., and Kherroubi Garcia, I, AI Fairness in Practice (2023 The Alan Turing Institute)

Journal Articles

Abercrombie G and others, 'A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms' (2024) arXiv preprint arXiv:2407.01294 https://doi.org/10.48550/arXiv.2407.01294 accessed 10 March 2026.

Alistair Reid, Simon O'Callaghan & Yaya Lu, Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources [2023] Gradient Institute & CSIRO

Aishat Oyenike Salami, Artificial intelligence, digital colonialism and the implications for Africa's future development [2024] Data for Policy Proceedings; AI, Ethics and Policy Governance in Africa

Alistair Reid, Simon O'Callaghan & Yaya Lu, Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources [2023] Gradient Institute & CSIRO

Amershi S and others, 'Guidelines for Human-AI Interaction' in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (ACM 2019) 1 https://doi.org/10.1145/3290605.3300233 accessed 10 March 2026.

Amy A. Winecoff & Elizabeth Anne Watkins, 'Artificial Concepts of Artificial Intelligence: Institutional Compliance and Resistance in AI Startups' [2022] Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society

Andrés Guadamuz, A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs [2024] GRUR International, Journal of European and International IP Law

Bagehorn F and others, 'AI Risk Atlas: Taxonomy and Tooling for Navigating AI Risks and Resources' (2025) arXiv preprint arXiv:2503.05780 https://arxiv.org/abs/2503.05780 accessed 10 March 2026.

Billiris G, Gill A and Bandara M, 'Privacy in the Age of AI: A Taxonomy of Data Risks' (2025) arXiv preprint arXiv:2510.02357 https://doi.org/10.48550/arXiv.2510.02357 accessed 10 March 2026.

Claudio Novelli, Mariarosaria Taddeo & Luciano Floridi, Accountability in Artificial Intelligence : What It Is and How It Works [2024] AI & SOCIETY Volume 39, issue 4

Delaram Golpayegani, Isabelle Hupont, Cecilla Panigutti, Harshvardhan J Pandit, Sven Schade, Declan O'Sullivan & Dave Lewis, AI Cards : Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act [2024] Computers and Society <arXiv:2406.18211>

Emmanouil Papagiannidis, Patrick Mikalef & Kieran Conboy, Responsible artificial intelligence governance : A review and research framework [2025] The Journal of Strategic Information Systems, Volume 34, Issue 2

Gebru T and others, 'Datasheets for Datasets' (2021) 64(12) Communications of the ACM 86 https://doi.org/10.1145/3458723 accessed 10 March 2026.

Hiniduma K, Bez JL, Madduri R and Byna S, 'AIDRIN: A Comprehensive Toolset for Automating Data Preparation for AI' (SC25 Research & ACM SRC Posters, 2025) https://sc25.conference-program.com/presentation/?id=int_post104&sess=sess539 accessed 10 March 2026.

Jacqueline Poucette, The Right to Be Forgotten and AI ; Legal Obligations, Technical Limits, and Innovation(Lund University 2025, Graduate Thesis) 31

Jahaziel Osei Mensah & Aimee Van Wynsberghe, Where are the missing values : an exploration of the need to incorporate Ubuntu values into African AI Policy [2025] AI and Ethics

Jana Gerlach, Paul Hoppe, Sarah Jagles, Luisa Licker & Micahel H. Bretiner, Decision support for efficient XAI services- A morphological analysis, business model archetypes, and a decision tree [2022] Electronic Markets

Joydeep Chandra & Satyam Kumar Navneet, Advancing Responsible Innovation in Agentic AI : A study of Ethical Frameworks for Household Automation [2025] arXiv (Cornell University)

Kars Alfrink, Ianus Keller, Geird Kortuem & Neelke Doorn, Contestable AI by Design : Towards a Framework [2022] Minds and Machines

Laura Waltersdorfer, Fajar J, Ekaputra, Tomasz Miksa & Marta Sabou, AuditMAI: Towards An Infrastructure for Continuous Auditing [2024] Computers and Society <arXiv:2406.14243>

Leila Methani, Andrea Aler Tubella, Virginia Dignum & Andreas Theodorou, Let Me Take Over: Variable Autonomy for Meaningful Human Control [2021] Frontiers in Artificial Intelligence , Volume 4

Liu X and others, 'AgentBench: Evaluating LLMs as Agents' (2023) arXiv preprint arXiv:2308.03688 https://doi.org/10.48550/arXiv.2308.03688 accessed 10 March 2026.

Marco Almada, Law & Compliance in AI Security & Data Protection; AI and Data Protection Training Module , European Data Protection Board [2024] <https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/law-compliance-ai-security-data_en>

Markus Kattnig, Allesa Angerschmid, Thomas Reichel & Roman Kern, Assessing trustworthy AI: Technical and legal perspectives of fairness in AI [2024] Computer Law& Security Review

Michelle Seng Ah Lee, Luciano Floridi & Jatinder Singh, Formalising trade-offs beyond algorithmic fairness : lessons from ethical philosophy and welfare economics [2021] AI and Ethics

McDermott P and others, Human-Machine Teaming Systems Engineering Guide (MITRE Product MP180941, MITRE Corporation 2018) https://www.mitre.org/sites/default/files/2021-11/prs-17-4208-human-machine-teaming-systems-engineering-guide.pdf accessed 10 March 2026.

Notice Pasipamire & Abton Muroyiwa, Navigating algorithm bias in AI: ensuring fairness and trust in Africa [2024] Front.Res.Metr, Anal, Vol 9

Olajide Babarunde Taofeek, Ekechi Chijioke Cyriacus, Popoola Taoheed Olawale, Adeshina Oguntoye Geroge , Ayittey Selasi & Ogueji Peter Chika Ozo, Machine learning for financial inclusion in agriculture : A study of AI-based credit scoring tools in rural Nigeria [2025] World Journal of Advanced Research and Reviews

Reuben Binns et al, 'Not Even Nice Work If You Can Get It; A Longitudinal Study of Uber's Algorithmic Pay and Pricing' in The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)

Roman AC and others, 'Open Datasheets: Machine-Readable Documentation for Open Datasets and Responsible AI Assessments' (2023) arXiv preprint arXiv:2312.06153 https://doi.org/10.48550/arXiv.2312.06153 accessed 10 March 2026.

Saeri AK and others, 'Mapping AI Risk Mitigations: Evidence Scan and Preliminary AI Risk Mitigation Taxonomy' (2025) arXiv preprint arXiv:2512.11931 https://doi.org/10.48550/arXiv.2512.11931 accessed 10 March 2026.

Siezte Kai Kuilman, Luciano Cavalcante Siebert, Stefan Buijsman & Catholjin M Jonker, How to gain control and influence algorithms : contesting AI to find relevant reasons [2024] AI and Ethics

Schwartz R and others, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (NIST Special Publication 1270, National Institute of Standards and Technology 2022) https://doi.org/10.6028/NIST.SP.1270 accessed 10 March 2026.

Shneiderman B, 'Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy' (2020) 36(6) International Journal of Human-Computer Interaction 495 https://doi.org/10.1080/10447318.2020.1741118 accessed 10 March 2026.

Stefan Schmager, Ilias O Pappas & Polyxeni Vassilakopoulo, Understanding Human-Centred AI: A review of its defining elements and a research agenda [2025] Behaviour & Information Technology , Volume 44, Issue 15

Stewart M and others, 'Datasheets for Machine Learning Sensors' (2023) arXiv preprint arXiv:2306.08848 https://doi.org/10.48550/arXiv.2306.08848 accessed 10 March 2026.

Swinger N and others, 'What Are the Biases in My Word Embedding?' in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (ACM 2019) 305 https://doi.org/10.1145/3306618.3314270 accessed 10 March 2026.

Online Resources

Citizenship Rights in Africa Initiative, Governing ID;Kenya's Huduma Namba Programme [2020] <https://digitalid.design/evaluation-framework-case-studies/kenya.html>

Constantine W, 'How to Detect and Mitigate Harmful Societal Bias in Your Organization's AI' (Xyonix, 19 October 2020) https://www.xyonix.com/blog/how-to-detect-and-mitigate-harmful-societal-bias-in-your-organizations-ai accessed 10 March 2026.

ISO/IEC 5338:2023, Information technology — Artificial intelligence — AI system life cycle processes https://www.iso.org/standard/81118.html accessed 10 March 2026.

Naadiya Moosajee , Fix AI's racist, sexist bias (14th March 2019) Mail & Guardian

<https://mg.co.za/article/2019-03-14-fix-ais-racist-sexist-bias/relevant local knowledge>

Lizzie Short, Building a Responsible AI Framework: 5 Key Principles for Organizations [2025] Harvard University: Professional & Executive Development

<https://professional.dce.harvard.edu/blog/building-a-responsible-ai-framework-5-key-principles-for-organizations/#What-Is-Responsible-AI >

Office of the Data Protection Commissioner, Guidance Note for Digital Credit Providers (December 2023) https://www.odpc.go.ke/wp-content/uploads/2024/02/ODPC-Guidance-Note-for-Digital-Credit-Providers.pdf accessed 10 March 2026.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689 https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng accessed 10 March 2026.

World Economic Forum, Why we need to care about Responsible AI in the age of the algorithm [2023]

< https://www.weforum.org/stories/2023/03/why-businesses-should-commit-to-responsible-ai/>

DIFA CONSULTANCY

UK International Development
Partnership | Progress | Prosperity

AKILI AI

KICTANet
The Power of Communities